

Aprendizagem e Extração do Conhecimento

PL02 · CRISP – DM | Data Understanding

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.

PARTE 01 · COMPREENSÃO DOS DADOS

CRISP-DM



Compreensão dos Dados

- **Recolha das fontes de dados relevantes**
- **Descrição: nº de registos, tipos de atributos, formato**
- **Exploração: estatística, visualizações, primeiras hipóteses**
- **Verificação da qualidade: falhas, inconsistências, outliers**

Compreensão dos Dados

- **Integridade referencial: algum id_paciente órfão em exames?**
- **Cobertura: há doentes sem qualquer exame?**
- **Cardinalidade: exames por doente, em média e máximo**
- **Um id órfão ou doente sem exames indica erro de recolha**

Compreensão dos Dados

Como escolher estatísticas e gráficos segundo o tipo de variável

Tipo	Exemplos nestes dados	Estatísticas adequadas
Numérica contínua	idade, imc, glicose_jejum, creatinina	média, mediana, desvio padrão, quartis
Numérica discreta	nº de exames por doente	moda, contagens, distribuição
Categórica nominal	sexo	frequências, moda
Binária	diabetico, fumador, sedentarismo	proporção, frequência
Temporal	data_visita	amplitude, ordenação, contagem por período

Compreensão dos Dados

- **Tendência central: média, mediana e moda**
- **Dispersão: desvio padrão, IQR, mínimo e máximo**
- **Forma: assimetria (enviesamento) e curtose**
- **Valores em falta: quantificar o padrão (tratar só na Data Preparation)**

Os problemas de qualidade a caçar

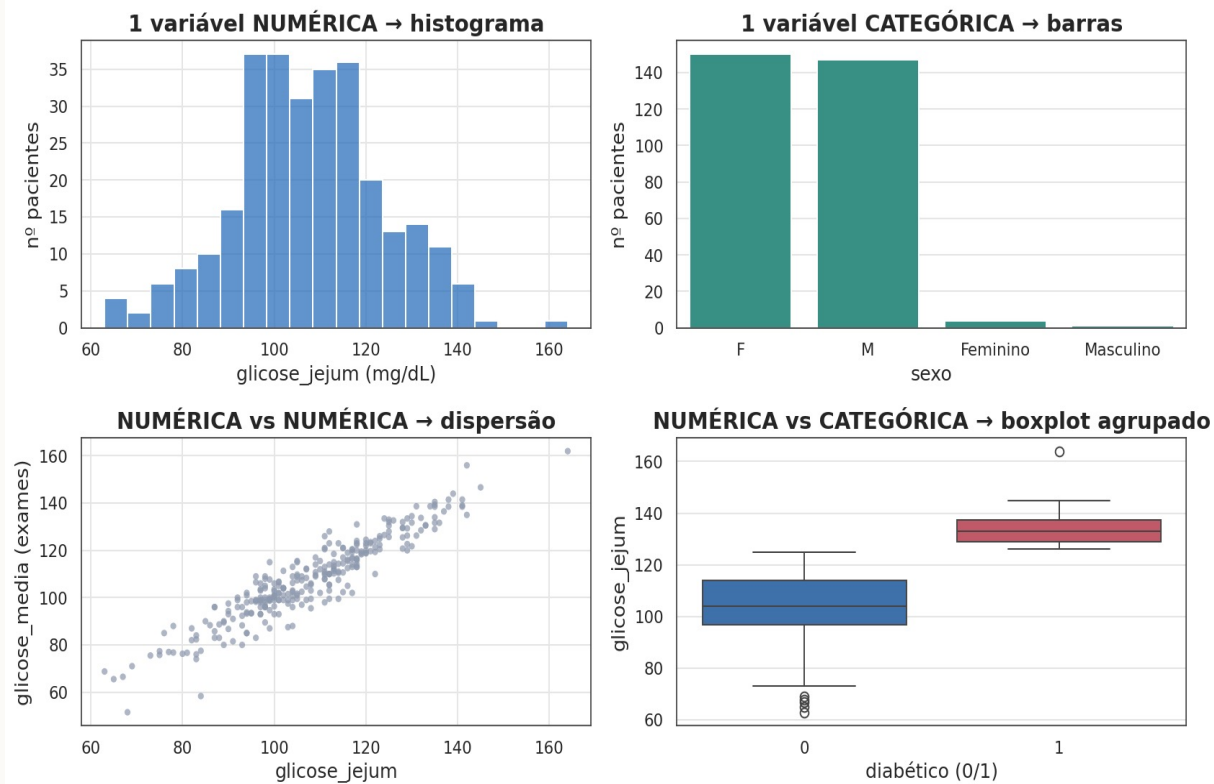
Cada problema tem uma forma simples de deteção na Data Understanding

Problema	Onde aparece nestes dados	Como se deteta
Duplicados	2 linhas repetidas em pacientes (P0004, P0051)	nº de linhas vs id únicos
Categorias inconsistentes	sexo: F/M mas também Feminino/Masculino	gráfico de barras
Outliers implausíveis	idade 137; imc 0 e 87.4; creatinina negativa; triglicéridos 1500	mínimo/máximo, boxplot
Valores em falta	colesterol_total (~10%), imc, hba1c...	% de nulos por coluna
Chave órfã	um id_paciente em exames sem par em pacientes	integridade referencial

Que gráfico usar?

Antes de resumir uma variável num número, olhe para a sua distribuição

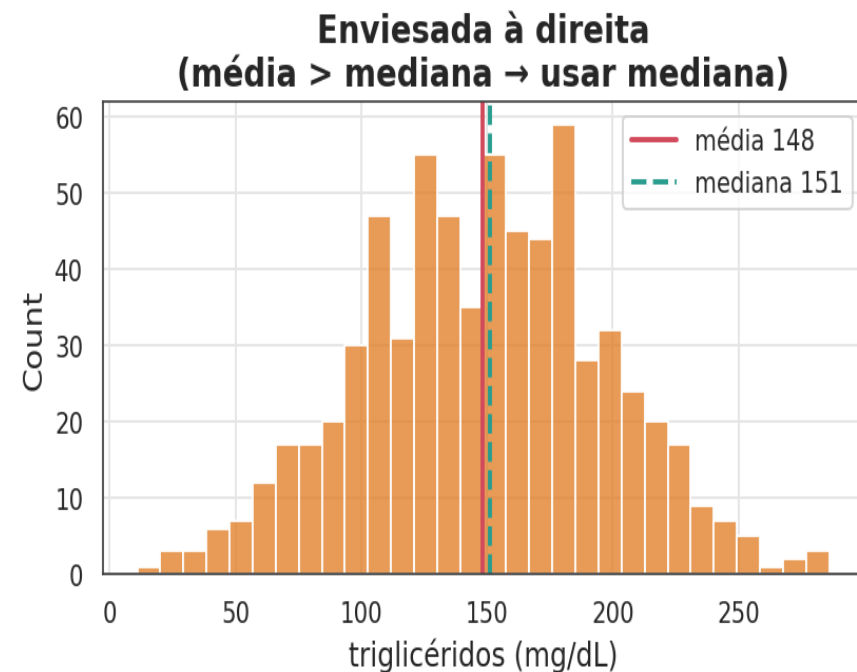
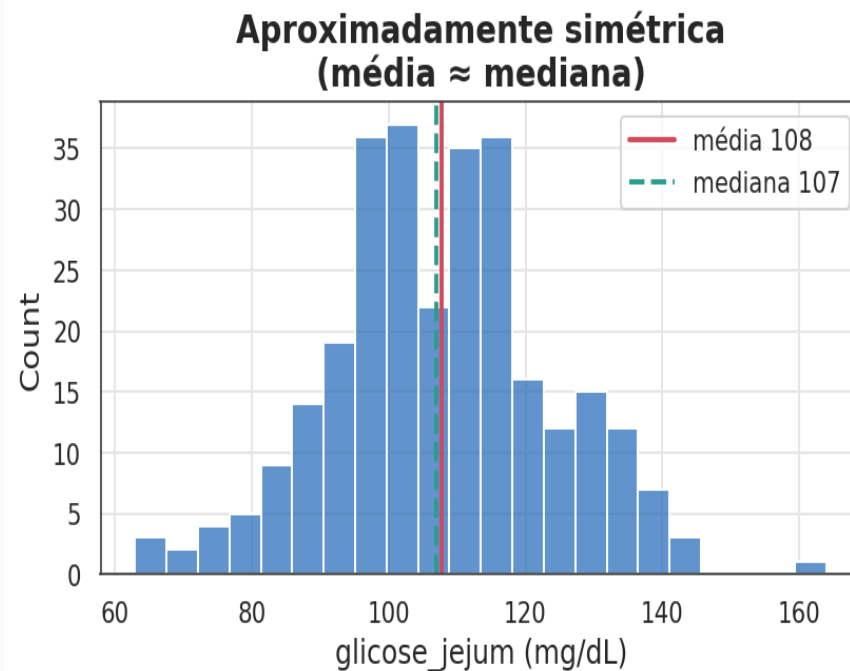
Que gráfico usar? Depende do tipo de variável



Histograma: a forma da distribuição

Simétrica → a média descreve bem; enviesada → a mediana é o valor típico

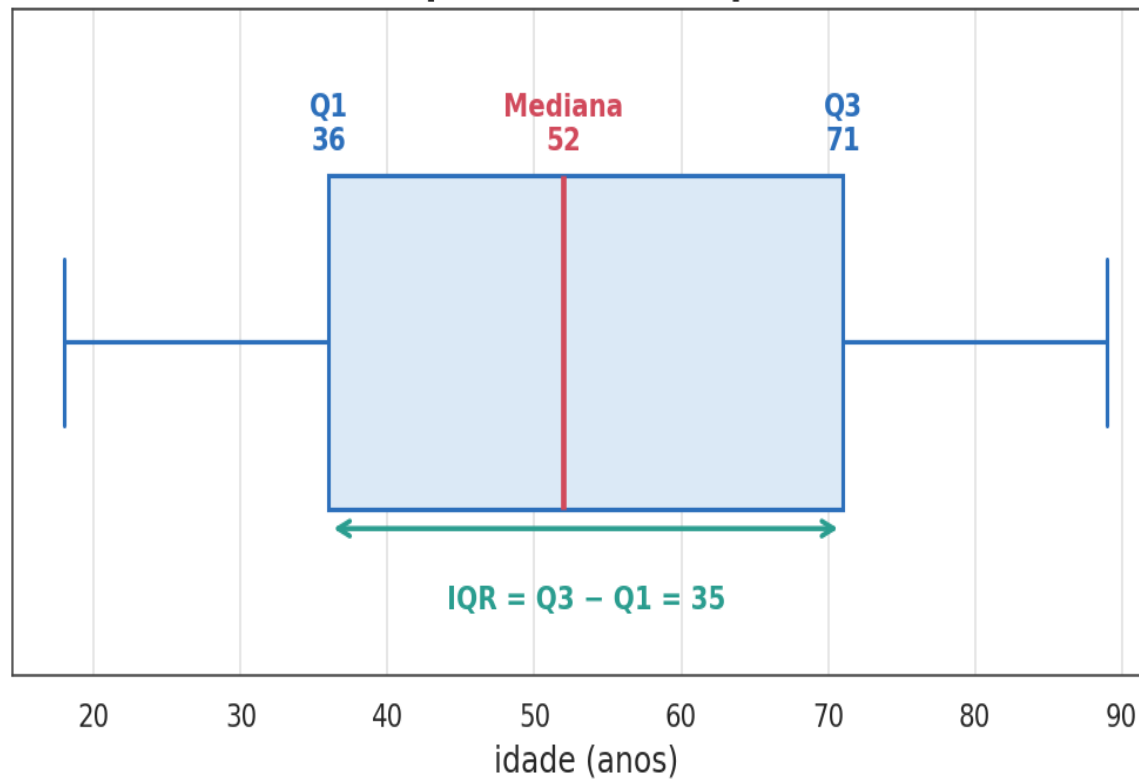
A forma da distribuição condiciona a estatística a usar



Anatomia de um boxplot

Num só gráfico: mediana, quartis, IQR e potenciais outliers

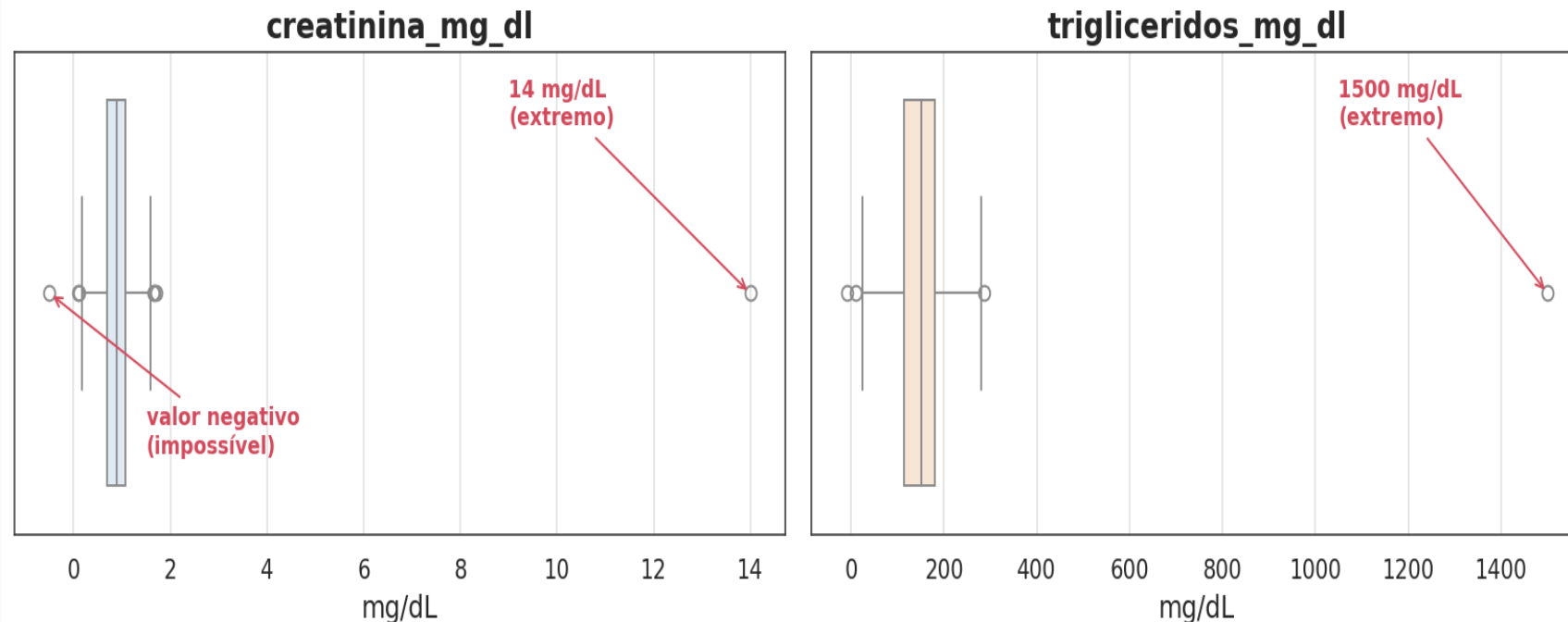
Anatomia de um boxplot: mediana, quartis, IQR e outliers



O boxplot revela valores implausíveis

Sem limpeza, uma creatinina negativa e triglicéridos de 1500 saltam à vista

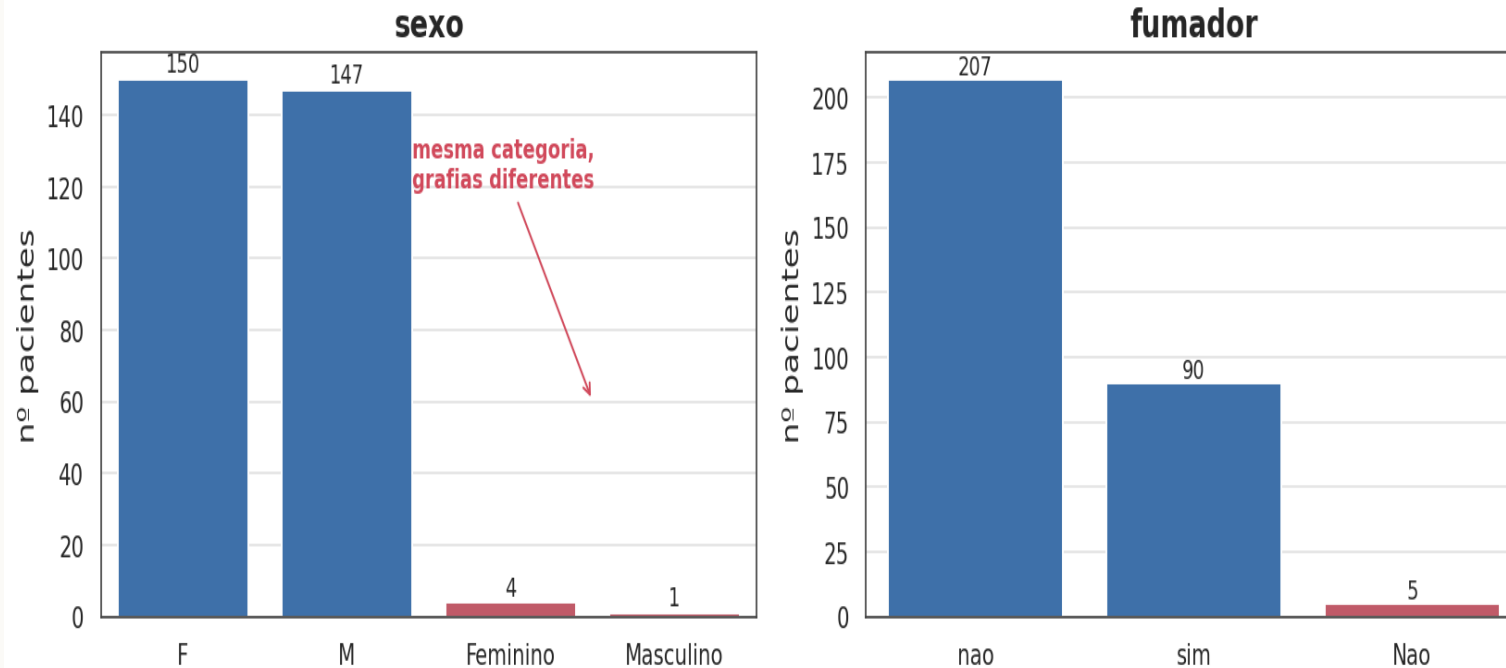
O boxplot revela valores implausíveis (dados por limpar)



Barras: categorias inconsistentes

Quatro categorias de sexo onde deviam existir apenas duas

Gráfico de barras expõe categorias inconsistentes



PARTE 02 • AI Studio

Compreensão dos Dados

Caso Diabetes

- Carregar e inspecionar os dados
- Identificar problemas de qualidade
- Explorar estatísticas e distribuições
- Visualizar relações entre atributos

AI Studio

O **AI Studio** é uma ferramenta comercial para análise de dados que utiliza machine learning e pode ser considerada uma alternativa para a ferramenta *Weka*.

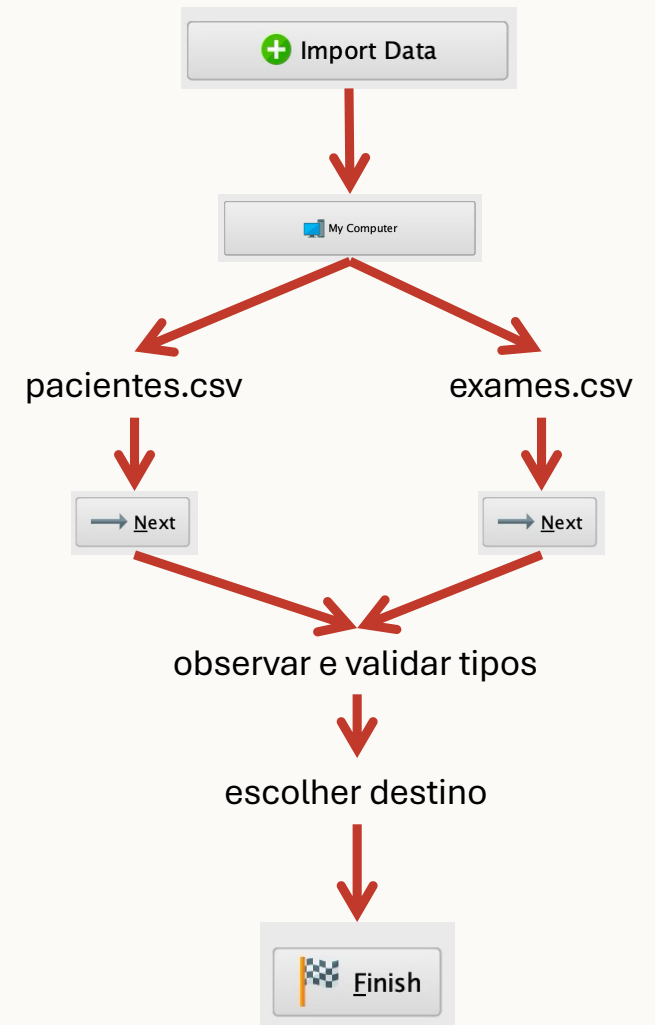
Esta ferramenta desenvolvida pela empresa com o mesmo nome, tem como principal missão acelerar o processo de criação de análises preditivas e torná-las mais fáceis para serem aplicadas em cenários práticos de negócios.



Compreensão dos Dados

Caso Diabetes

- **Carregar e inspecionar os dados**
- Identificar problemas de qualidade
- Explorar estatísticas e distribuições
- Visualizar relações entre atributos



Compreensão dos Dados

Caso Diabetes - pacientes.csv

id_paciente: identificador único do paciente

idade: idade do paciente, em anos

sexo: sexo do paciente

imc: índice de massa corporal

tensao_sistolica: tensão arterial sistólica, em mmHg

tensao_diastolica: tensão arterial diastólica, em mmHg

colesterol_total: colesterol total, em mg/dL

fumador: indica se o paciente é fumador

sedentarismo: indica se o paciente tem um estilo de vida sedentário

historico_familiar_diabetes: indica se há antecedentes familiares de diabetes

glicose_jejum: glicose em jejum, em mg/dL

diabetico: atributo-alvo; indica se o paciente tem diabetes

Compreensão dos Dados

Caso Diabetes - exames.csv

id_exame: identificador único do exame

id_paciente: chave que liga o exame ao paciente em *pacientes.csv*

data_visita: data em que o exame foi realizado

glicose_mg_dl: glicose medida na visita, em mg/dL

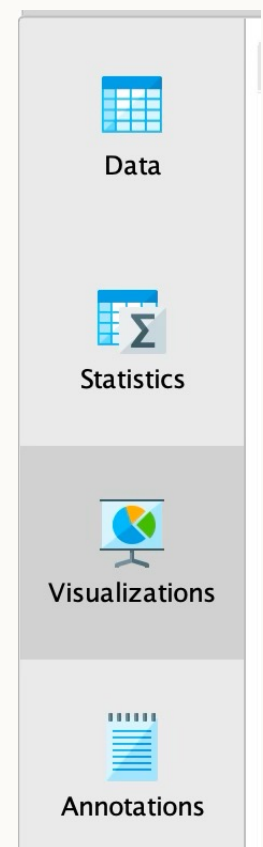
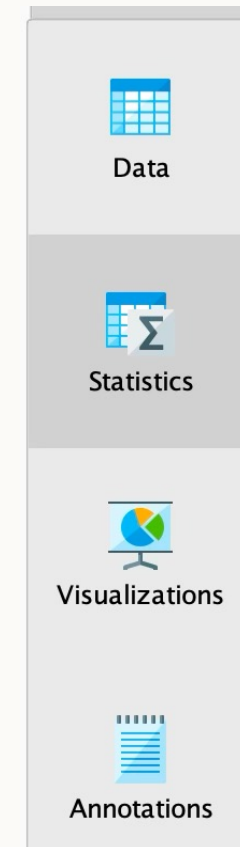
hba1c_percent: hemoglobina glicada, em percentagem; reflete o controlo da glicose nos últimos meses

creatinina_mg_dl: creatinina, em mg/dL; indicador da função renal

trigliceridos_mg_dl: triglicéridos, em mg/dL

Compreensão dos Dados: Caso Diabetes

- Carregar e inspecionar os dados
- **Identificar problemas de qualidade**
- **Explorar estatísticas e distribuições**
- **Visualizar relações entre atributos**



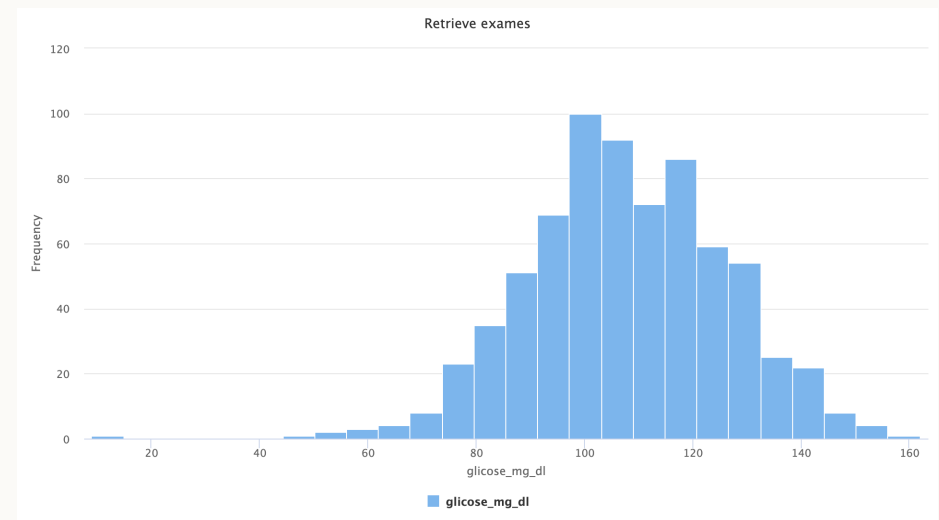
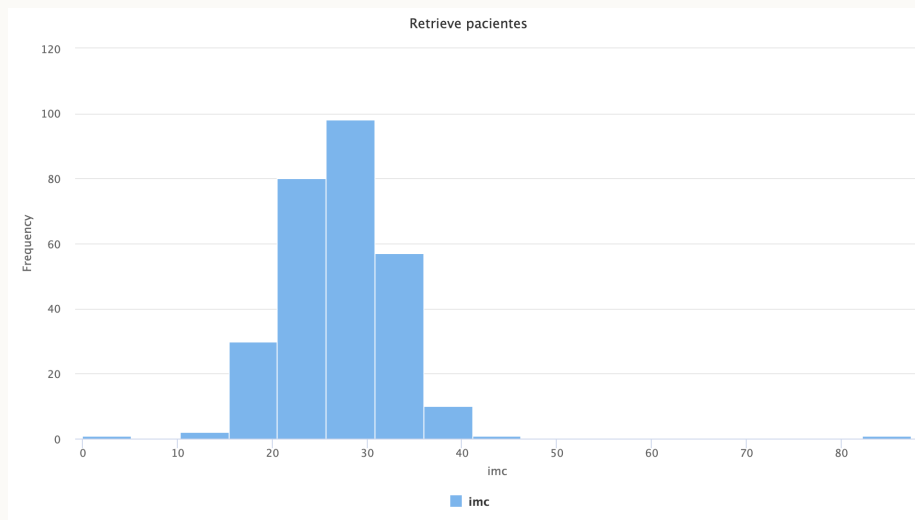
Metadados e Qualidade

Avaliar os metadados dos dados importados para os 2 datasets!

Name	Type	Missing	Statistics		
✓ id_paciente	Nominal	0	Least P0300 (1)	Most P0004 (2)	Values P0004 (2), P0051 (2), ...[298 more]
✓ idade	Integer	0	Min 18	Max 137	Average 53.586
✓ sexo	Nominal	0	Least Masculino (1)	Most F (150)	Values F (150), M (147), ...[2 more]
✓ imc	Real	22	Min 0	Max 87.400	Average 27.177
✓ tensao_sistolica	Real	0	Min 75	Max 280	Average 126.974
✓ tensao_diastolica	Real	15	Min 10	Max 117	Average 81.411
✓ colesterol_total	Real	30	Min 9	Max 294	Average 199.426
✓ fumador	Nominal	0	Least Nao (5)	Most nao (207)	Values nao (207), sim (90), ...[1 more]
✓ sedentarismo	Nominal	0	Least sim (124)	Most nao (178)	Values nao (178), sim (124)
✓ historico_familiar_diabetes	Nominal	9	Least sim (78)	Most nao (215)	Values nao (215), sim (78)
✓ glicose_jejum	Real	12	Min 63	Max 164	Average 107.634
✓ diabetico	Integer	0	Min 0	Max 1	Average 0.129

Visualizações – pacientes.csv

Visualizar **histogramas** dos diferentes atributos. Tentar perceber tendências



Visualizações – Classe ?

Gráfico de barras de uma potencial classe: Diabético

Plot 1 

Plot type

 Bar (Column) ▼

Value column

diabetico ▼

☒ Aggregate data

Group by

- ▼

Aggregation Function

Count ▼

Sorting

None ▼

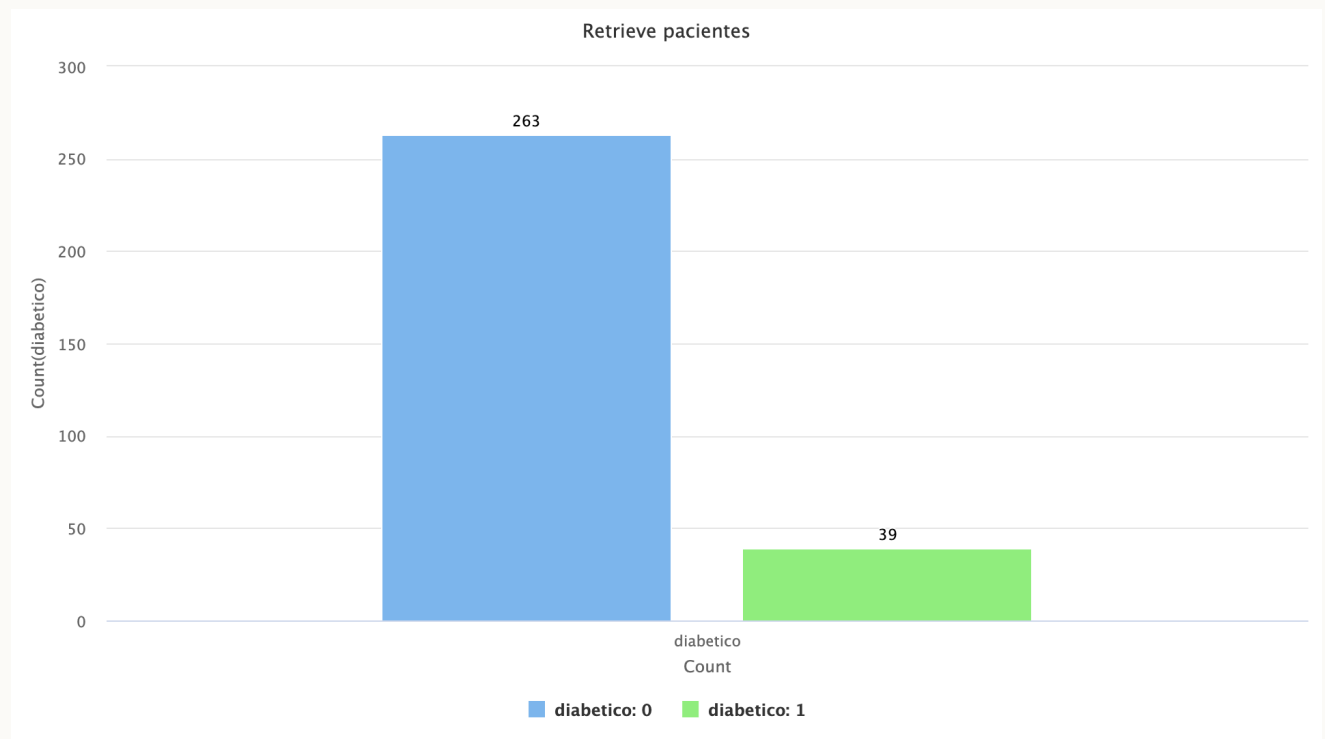
Color Group

diabetico ▼

Stacking

No stacking ▼

Plot style >>



Visualizações – Insights??

Criar um Scatter Plot com glicose_jejum no eixo X e imc no eixo Y, com cor por diabético.

Plot 1 

Plot type

 Scatter / Bubble ▼

X-Axis column

glicose_jejum ▼

Value column

imc ▼

Color

diabetico ▼

Size

– ▼

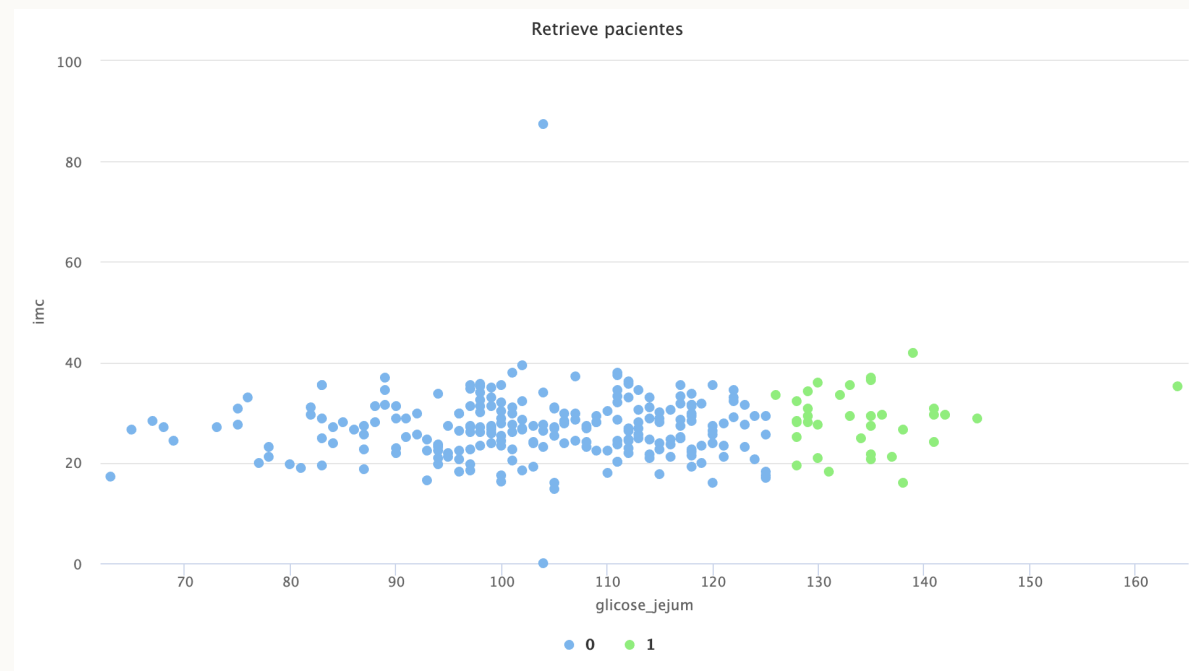
Jitter

 ☐

Regression interpolation

None ▼

Plot style >>



Visualizações – Insights??

Criar um Scatter Plot com glicose_jejum no eixo X e idade no eixo Y, com cor por diabético.

Plot 1 

Plot type

 Scatter / Bubble ▼

X-Axis column

glicose_jejum ▼

Value column

idade ▼

Color

diabetico ▼

Size

– ▼

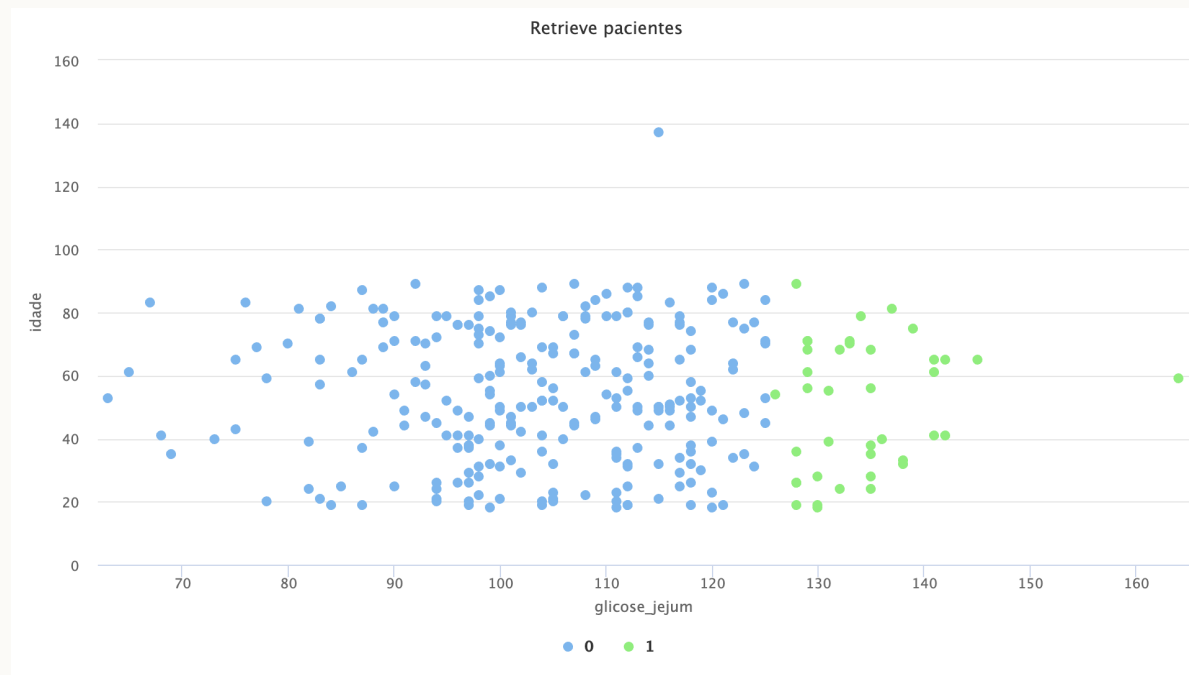
Jitter

 ☐

Regression interpolation

None ▼

Plot style >>



Exercício Conjunto

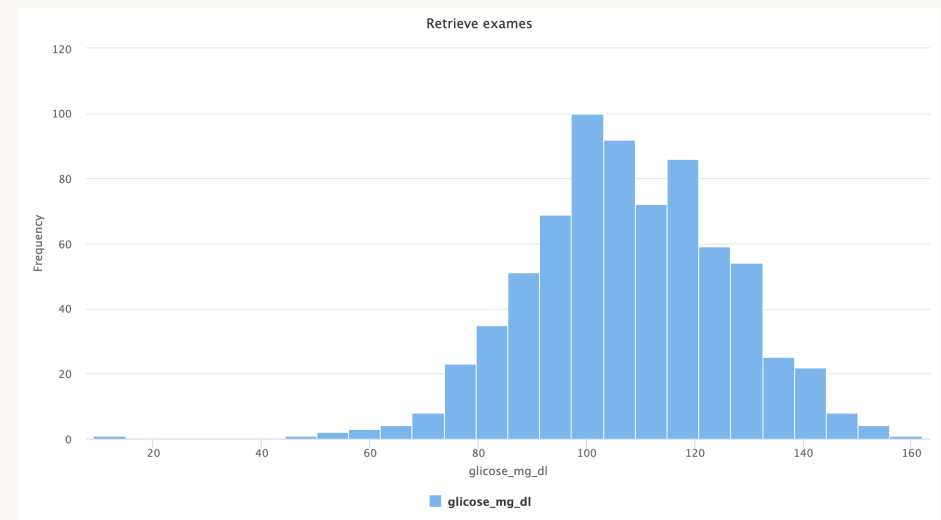
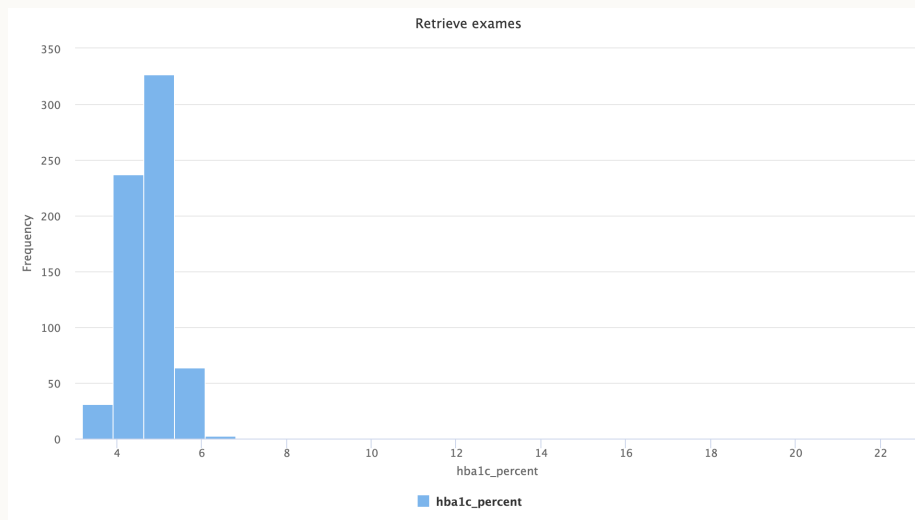
Compreensão dos Dados:

Os pacientes diabéticos aparecem mais concentrados em alguma zona do gráfico?



Visualizações – exames.csv

Criar um Histograma para hba1c_percent e outro para glicose_mg_dl.



Qualidade dos Dados

- Reparar que *pacientes.csv* tem 1 linha por paciente, mas *exames.csv* tem entre 1 e 4 linhas por paciente.
- Esta diferença de granularidade terá de ser resolvida no Data Preparation (ex.: Aggregate antes do Join).
- Verificar se existe algum id_paciente em *pacientes.csv* sem nenhuma linha correspondente em *exames.csv*.

Compreensão dos dados

- **Dois datasets, granularidades diferentes**
- **Valores em falta em ambos os ficheiros**
- **Forte desequilíbrio de classes em diabético**
- **Primeiras pistas visuais: glicose_jejum parece distinguir bem os dois grupos**

PARTE 03 · Python

Preparar o ambiente: VSCode + venv

Um ambiente virtual isolado, só para esta aula

1. abrir o vscode e o terminal:

2. criar a pasta da aula:

```
base > mkdir aec-python  
base > cd aec-python
```

3. criar um ambiente virtual novo

```
base > python -m venv .p102
```

4. ativar o ambiente

```
base > source .p102/bin/activate
```

Instalar os pacotes

pandas, seaborn, matplotlib e jupyter dentro do venv

5. com o novo venv (.pl02) ativo, instalar os pacotes

```
.pl02 > pip install pandas seaborn matplotlib jupyter
```

6. confirmar a instalação

```
.pl02 > python -c "import pandas, seaborn; print('ok')"
```

7. instalar a extensão do Jupyter e criar um ficheiro "aec-pl02.ipynb"

8 . Selecionar o interpretador:

Ctrl+Shift+P -> Python: Select Interpreter

Selecionar o .pl02

Ler os dados

Duas linhas para carregar os ficheiros e ver dimensões e tipos

```
import pandas as pd
import seaborn as sns

# ler os dois ficheiros
pac = pd.read_csv('pacientes.csv')
exa = pd.read_csv('exames.csv')

print(pac.shape, exa.shape)
print(pac.dtypes)
```

Estatística descritiva

`describe()` dá o mesmo que o operador `Statistics` (repare no máximo de idade)

estatística descritiva das variáveis numéricas
`pac.describe().round(1)`

	idade	imc	tensao_sistolica	tensao_diastolica	colesterol_total	glicose_jejum	diabetico
count	302.0	280.0	302.0	287.0	272.0	290.0	302.0
mean	53.6	27.2	127.0	81.4	199.4	107.6	0.1
std	21.7	6.6	21.1	13.3	37.0	16.2	0.3
min	18.0	0.0	75.0	10.0	9.0	63.0	0.0
25%	36.0	23.3	113.0	73.0	176.8	97.2	0.0
50%	53.0	27.2	125.5	81.0	201.0	107.0	0.0
75%	71.0	30.8	139.0	91.0	225.0	118.0	0.0
max	137.0	87.4	280.0	117.0	294.0	164.0	1.0

Valores em falta

A percentagem de nulos por coluna, numa só linha

valores em falta, em %

```
(pac.isna().mean() * 100).round(1)
```

```
id_paciente      0.0
idade            0.0
sexo             0.0
imc              7.3
tensao_sistolica 0.0
tensao_diastolica 5.0
colesterol_total  9.9
fumador          0.0
sedentarismo     0.0
historico_familiar_diabetes 3.0
glicose_jejum    4.0
diabetico        0.0
dtype: float64
```

Categorias e duplicados

value_counts expõe o sexo mal codificado; há duas linhas repetidas

categorias distintas

```
pac["sexo"].value_counts()
```

```
sexo
F          150
M          147
Feminino    4
Masculino    1
Name: count, dtype: int64
```

Relacionar as duas tabelas

Detetar o id órfão, agregar por doente e juntar com merge

```
# exames de doentes inexistentes
t = set(exa["id_paciente"]) - set(pac["id_paciente"])
print(t)

# glicose média por doente e juntar
g = (exa.groupby("id_paciente")["glicose_mg_dl"].mean())
print(g)

# Quem não tem exames fica com NaN na glicose
df = pac.merge(g, on="id_paciente", how="left")
print(df)
```

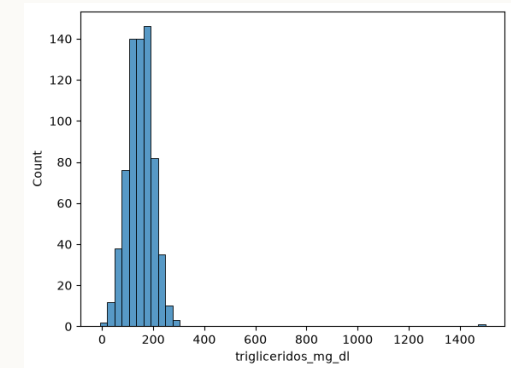
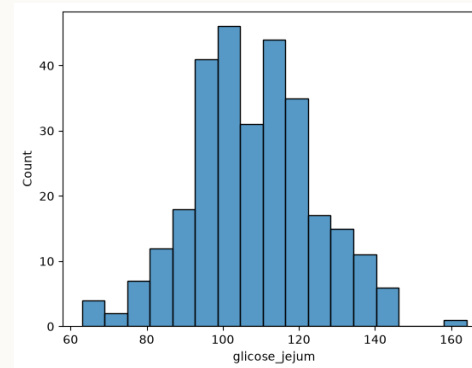
Histograma

A mesma leitura de forma e enviesamento, agora em seaborn

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(pac["glicose_jejum"])
plt.show()

sns.histplot(exa["trigliceridos_mg_dl"])
plt.show()
```



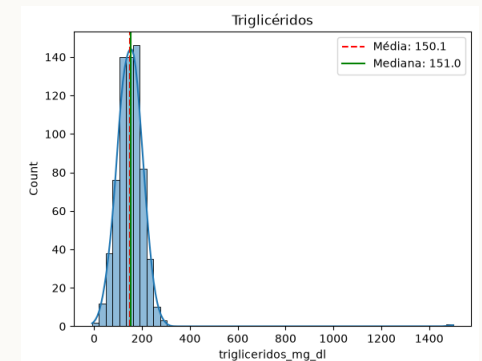
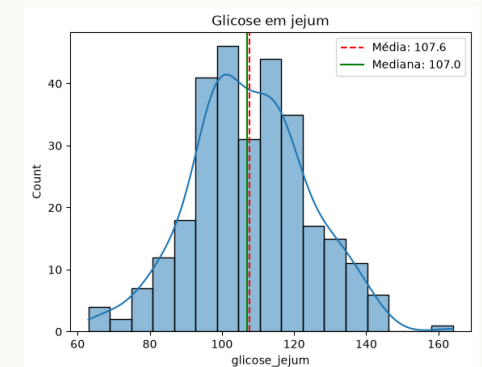
Histograma

A mesma leitura de forma e enviesamento, agora em seaborn

```
def hist_com_estatisticas(serie, titulo=None):  
    media = serie.mean()  
    mediana = serie.median()
```

```
    sns.histplot(serie, kde=True)  
    plt.axvline(media, color="red", linestyle="--", label=f"Média: {media:.1f}")  
    plt.axvline(mediana, color="green", linestyle="-", label=f"Mediana: {mediana:.1f}")  
    plt.title(titulo)  
    plt.legend()  
    plt.show()
```

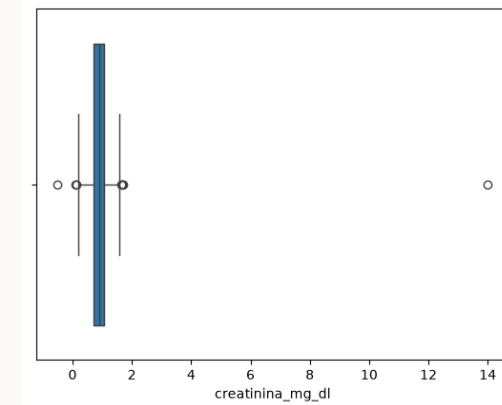
```
hist_com_estatisticas(pac["glicose_jejum"], "Glicose em jejum")  
hist_com_estatisticas(exa["trigliceridos_mg_dl"], "Triglicéridos")
```



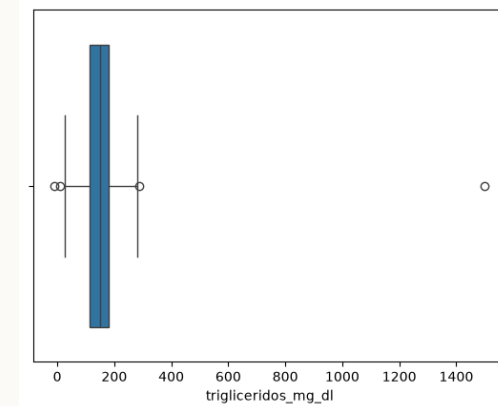
Boxplot: valores implausíveis

Sem limpeza, os valores impossíveis saltam à vista

```
sns.boxplot(x=exa["creatinina_mg_dl"])
plt.show()
```



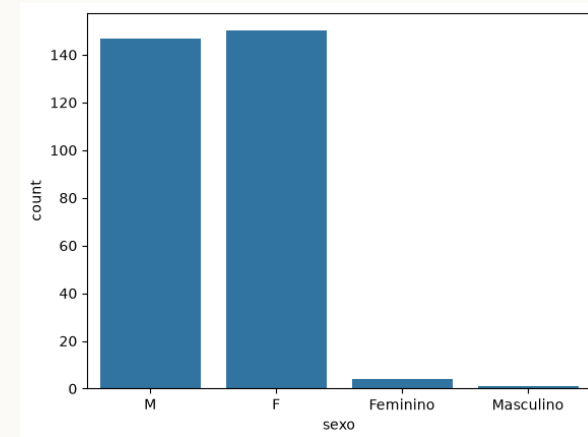
```
sns.boxplot(x=exa["trigliceridos_mg_dl"])
plt.show()
```



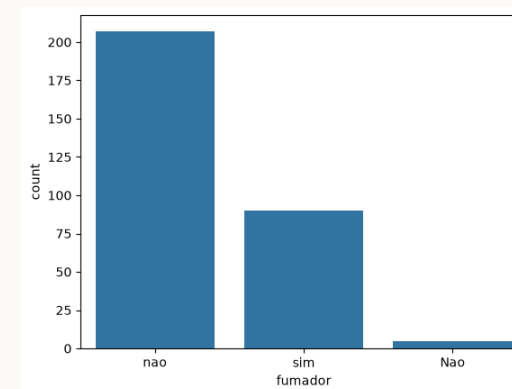
Barras: categorias inconsistentes

countplot revela as categorias a uniformizar

```
sns.countplot(data=pac, x="sexo")
plt.show()
```



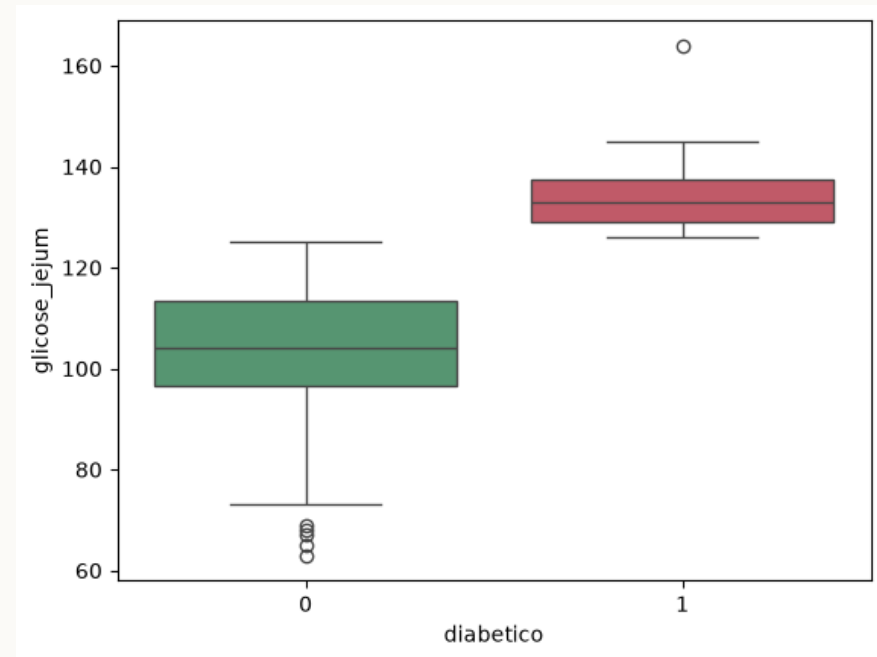
```
sns.countplot(data=pac, x="fumador")
plt.show()
```



Comparar grupos

boxplot de glicose por diabético: a glicose separa bem os grupos

```
sns.boxplot(  
    data=pac,  
    x="diabetico",  
    y="glicose_jejum",  
    hue="diabetico",  
    palette={0: "#4C9F70", 1: "#D1495B"},  
    legend=False,  
)  
plt.show()
```



AI Studio vs Python

AI Studio

- exploração visual muito rápida
- arrastar e ligar operadores
- difícil de repetir e versionar
- poucos passos ficam registados

Python

- controlo total e fino
- reproduzível e versionável (git)
- fácil de repetir noutro dataset
- o notebook documenta o processo

O raciocínio estatístico é o mesmo; muda a rapidez vs a reprodutibilidade.

Ficha de Exercícios 02

FE02 · Data Understanding em Python

- Ler pacientes.csv e exames.csv com pandas
- Reproduzir describe(), valores em falta e categorias
- Relacionar as tabelas (groupby + merge) e visualizar
- Comparar a experiência com o AI Studio

Aprendizagem e Extração do Conhecimento

PL02 · CRISP – DM | Data Understanding

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.