Análise de Dados - Exercicio 2

Aluno: Cláudio Mendes Ramos Número: 2022121538

Características do Dataset

Fonte: https://www.kaggle.com/datasets/ak0212/uae-cancer-patient-dataset

Registos: 10.000 pacientes
Atributos: 20 colunas
Dados demográficos e clínicos: inclui idade, género, tipo de cancro, tipo de tratamento, resultados clínicos, etc.

Observações-Chave

Valores em Falta

Apenas uma coluna apresenta valores em falta:

death_date: 9.008 valores ausentes
↳ Isto é esperado, já que a maioria dos pacientes ainda está viva.

Outliers

Foram analisadas as colunas age, weight e height com base na regra do IQR (Intervalo Interquartílico):

age:
- Limites: -18 a 126
- Outliers: 0

weight:
- Limites: 30 a 110
- Outliers: 64

height:
- Limites: 143.5 cm a 195.5 cm
- Outliers: 95

Todos os outliers já foram removidos ou tratados antes da exportação deste ficheiro.

Resumo das Estratégias de Limpeza

Valores em Falta

- comorbidities: valores ausentes substituídos por 'none'
- cause_of_death: valores ausentes sustituidos por 'alive'
- death_date: mantido com valores nulos, pois indica naturalmente que o paciente não faleceu

Impacto da Limpeza

Redução do total de linhas de 10.000 para 9.841 (159 registos com outliers foram eliminados).
Os dados estão agora limpos e prontos para análise estatística ou modelagem.