

Mineração de Dados

Tratamento de Outliers

Ettore Motta Gazzinelli

# Sumário

<b>1</b>	<b>Análise Exploratória de Outliers</b>	<b>2</b>
1.1	Análise Categórica (Frequência) . . . . .	2
1.2	Análise Numérica (Boxplots) . . . . .	4
<b>2</b>	<b>Metodologia de Tratamento</b>	<b>7</b>
2.1	Decisão sobre Outliers Categóricos vs. Numéricos . . . . .	7
2.2	Transformação Logarítmica (Variáveis Numéricas) . . . . .	7
2.3	Discretização da Variável IMC . . . . .	8

# Capítulo 1

## Análise Exploratória de Outliers

A análise de outliers foi dividida em duas frentes: a análise de variáveis categóricas, buscando por frequências raras, e a análise de variáveis numéricas, buscando por valores estatisticamente extremos.

### 1.1 Análise Categórica (Frequência)

Para as variáveis categóricas, investigamos a frequência de cada classe. "Outliers" neste contexto são classes com uma representatividade muito baixa (ex:  $< 1\%$ ), que poderiam ser erros de digitação ou eventos genuinamente raros.

Abaixo estão as tabelas de frequência para as variáveis-alvo (doenças) e um exemplo de variável de perfil (P01101).

**Tabela 1.1:** Contagem de valores para a coluna: 'P01101'

Valor (P01101)	Contagem	Porcentagem (%)
3	20027	22.57
2	17542	19.77
7	12241	13.79
4	10716	12.08
1	9345	10.53
5	8481	9.56
0	7348	8.28
6	3036	3.42

**Tabela 1.2:** Contagem de valores para a coluna: 'infarto'

Valor (infarto)	Contagem	Porcentagem (%)
2 (Não)	87356	98.44
1 (Sim)	1380	1.56

**Tabela 1.3:** Contagem de valores para a coluna: 'angina'

Valor (angina)	Contagem	Porcentagem (%)
2 (Não)	88060	99.24
1 (Sim)	676	0.76

**Tabela 1.4:** Contagem de valores para a coluna: 'insuficiencia\_cardiaca'

Valor	Contagem	Porcentagem (%)
2 (Não)	87711	98.84
1 (Sim)	1025	1.16

**Tabela 1.5:** Contagem de valores para a coluna: 'arritmia'

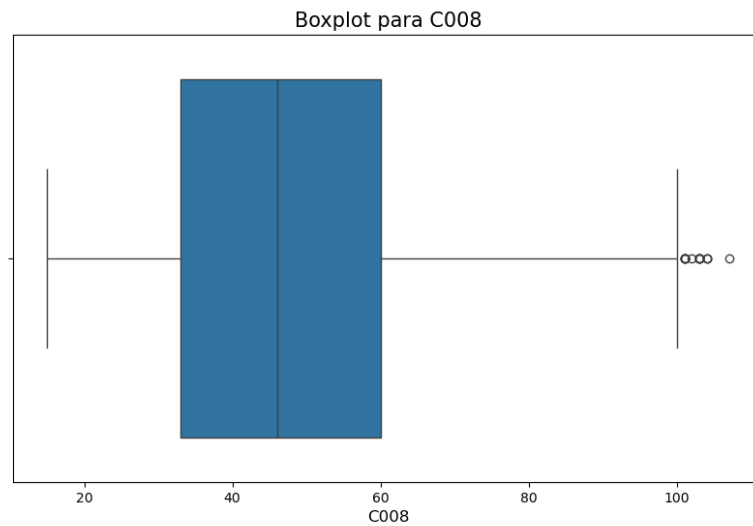
Valor (arritmia)	Contagem	Porcentagem (%)
2 (Não)	86663	97.66
1 (Sim)	2073	2.34

**Tabela 1.6:** Contagem de valores para a coluna: 'outra\_doenca\_cardiaca'

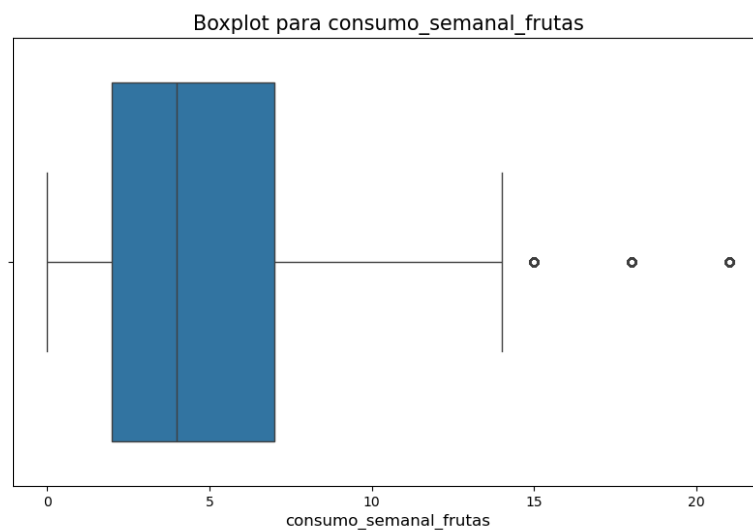
Valor	Contagem	Porcentagem (%)
2 (Não)	87912	99.07
1 (Sim)	824	0.93

## 1.2 Análise Numérica (Boxplots)

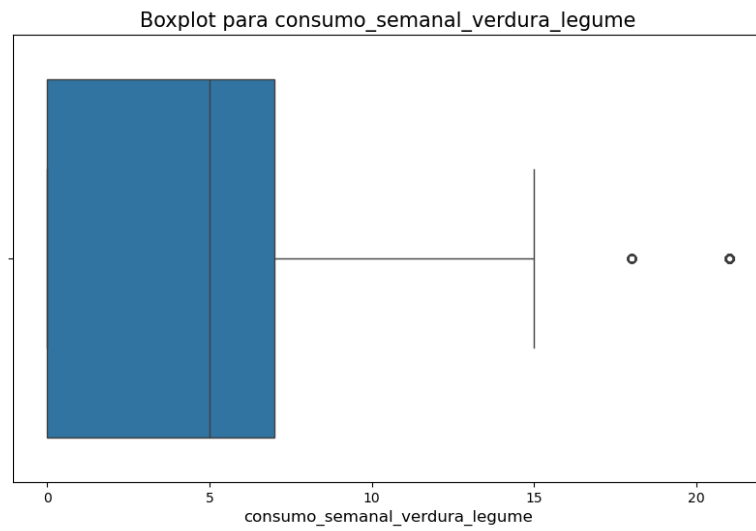
Para as variáveis numéricas, utilizamos boxplots para identificar visualmente os outliers. Estes são definidos como pontos de dados que caem fora do intervalo de  $1.5 \times$  o Intervalo Interquartil (IQR).



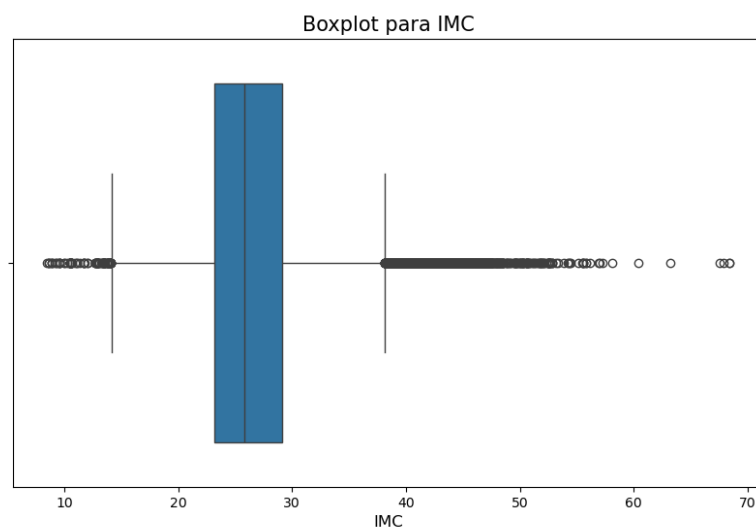
**Figura 1.1:** Boxplot para C008 (Idade)



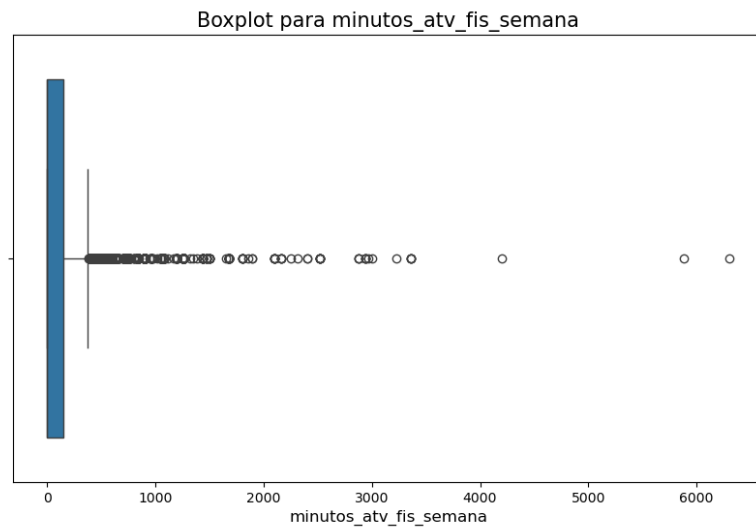
**Figura 1.2:** Boxplot para Consumo Semanal de Frutas



**Figura 1.3:** Boxplot para Consumo Semanal de Verduras/Legumes



**Figura 1.4:** Boxplot para IMC



**Figura 1.5:** Boxplot para Minutos de Atividade Física por Semana

## Capítulo 2

# Metodologia de Tratamento

### 2.1 Decisão sobre Outliers Categóricos vs. Numéricos

Após a análise exploratória, foi decidido **não remover** os outliers.

Para os dados **categóricos**, as classes com baixa frequência (como "Sim" para 'angina', com 0.76%) não são erros, mas sim o evento-alvo de estudo. Removê-los eliminaria a informação crucial para o modelo.

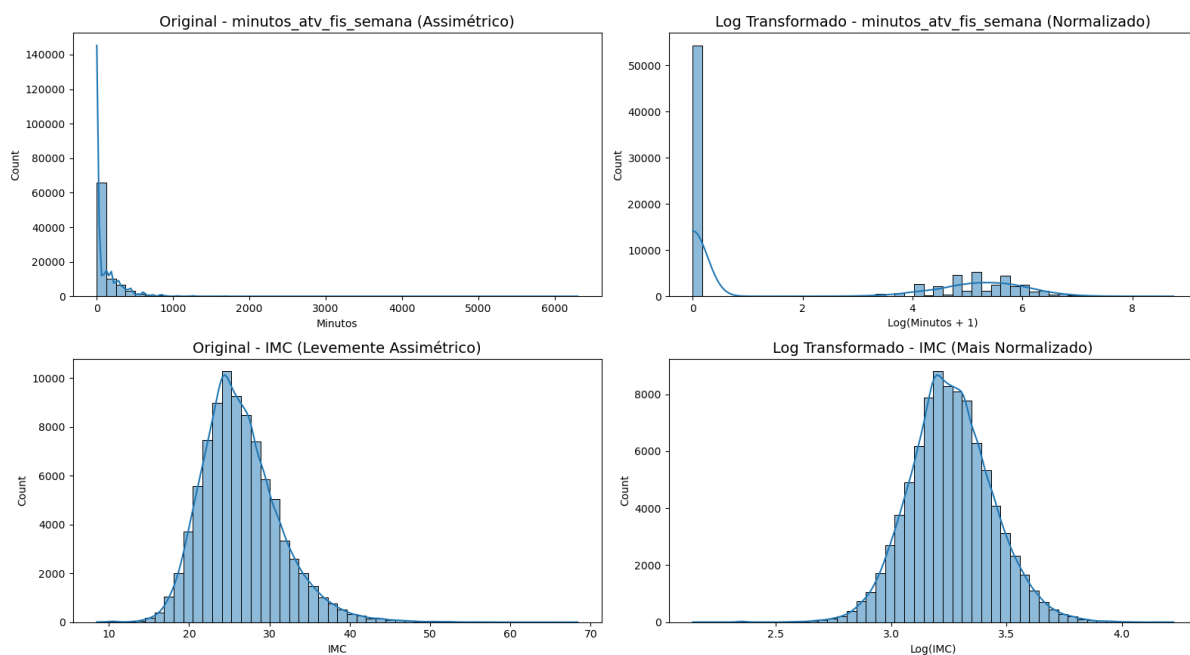
Para os dados **numéricos** (como 'IMC' e 'minutos\_atv\_fis\_semana'), os valores extremos (ex: obesidade ou pessoas muito ativas) são parte válida da população e essenciais para a generalização do modelo.

Portanto, a estratégia de tratamento não foi a remoção, mas a **transformação** e **discretização**.

### 2.2 Transformação Logarítmica (Variáveis Numéricas)

Para reduzir a assimetria (skewness) causada pelos valores extremos em 'minutos\_atv\_fis\_semana' e 'IMC' e normalizar sua distribuição, foi aplicada a transformação logarítmica (log1p). A Figura 2.1 demonstra o "antes e depois", mostrando como a transformação aproxima os dados de uma distribuição normal.

### Comparação: Original vs. Transformação Logarítmica



**Figura 2.1:** Comparação: Distribuição Original vs. Log Transformada

## 2.3 Discretização da Variável IMC

Embora a transformação logarítmica seja estatisticamente útil, para a variável 'IMC', a **discretização** (ou **binning**) em categorias clinicamente relevantes (baseadas na OMS) preserva melhor o significado dos dados.

Esta abordagem transforma o 'outlier' (ex: IMC 45) em uma categoria ('Obesidade Grau III'), que é mais interpretável e robusta para modelos de árvore de decisão. A Tabela 2.1 mostra o resultado dessa discretização.

**Tabela 2.1:** Distribuição de Frequência da nova coluna 'IMC\_Categoria'

Categoria (IMC)	Contagem
Abaixo do Peso	2247
Peso Normal	34386
Sobrepeso	33555
Obesidade Grau I	13635
Obesidade Grau II	3743
Obesidade Grau III	1170