1. Introdução e objetivo

Este relatório tem como objetivo realizar uma análise exploratória inicial do **UCI Energy Efficiency Dataset**, com foco na compreensão de suas principais características e na identificação de padrões relevantes que possam orientar o desenvolvimento de modelos preditivos para estimativa do consumo energético de aquecimento (Heating Load) e resfriamento (Cooling Load).

A estrutura do documento inclui a descrição do conjunto de dados, seguida por análises univariadas e multivariadas, com o intuito de extrair insights que embasem decisões futuras na etapa de modelagem preditiva.

---

2. Descrição do Conjunto de Dados

- **Amostras**: 768 edifícios simulados em clima mediterrâneo.
- **Variáveis de Entrada (8):**
    - Relative Compactness
    - Surface Area (m²)
    - Wall Area (m²)
    - Roof Area (m²)
    - Overall Height (m)
    - Orientation (1–4)
    - Glazing Area (fração da área de fachada)
    - Glazing Area Distribution (1–5)
- **Variáveis de Saída (2):**
    - Heating Load (kWh/m²)
    - Cooling Load (kWh/m²)

Não há valores faltantes no dataset: todas as 768 amostras estão completas.

---

3. Estatísticas Descritivas Gerais

Variável	Mínimo	Média	Mediana	Máximo	Desvio Padrão
Relative Compactness	0.62	0.81	0.82	0.98	0.10
Surface Area (m²)	514	645	647	808	84
Wall Area (m²)	245	294	296	416	48
Roof Area (m²)	110	151	151	220	30
Overall Height (m)	3.5	4.75	4.75	7.0	1.06
Glazing Area (fração)	0.00	0.10	0.06	0.40	0.12
Heating Load (kWh/m²)	6.02	24.70	24.30	43.00	7.49
Cooling Load (kWh/m²)	10.50	21.10	20.10	43.20	5.84

Observação: As variáveis contínuas apresentam variabilidade moderada; as variáveis categóricas (Orientation, Glazing Area Distribution) devem ser tratadas como dummy na modelagem.

---

4. Análise Univariada

1. **Relative Compactness**
    - Distribuição ligeiramente enviesada à esquerda (mais edifícios com alta compacidade).
    - Valor modal em torno de 0.82, indicando preferência por projetos compactos.
2. **Surface Area, Wall Area e Roof Area**
    - Todas apresentam distribuição aproximadamente normal, centradas nos valores médios apresentados.
    - A correlação intrínseca entre essas áreas é esperada: edifícios maiores tendem a ter todas as áreas aumentadas.
3. **Overall Height**
    - Distribuição bimodal em 3.5 m e 7.0 m, refletindo projetos de 1 e 2 pavimentos.

4. **Glazing Area**
    - Muitos edifícios sem aberturas (0.0) e alguns com proporções médias (0.125, 0.25, 0.375, 0.5).
    - Média baixa (0.10), indicando que a maioria das fachadas tem pouca janela.
5. **Saídas (Heating e Cooling Load)**
    - Heating Load distribui-se aproximadamente normal, média em 24.7 kWh/m².
    - Cooling Load tem cauda direita mais pronunciada, indicando alguns edifícios com alta demanda de resfriamento.

---

5. Análise de Correlações

- **Matriz de Correlação:**
    - Heating Load & Relative Compactness: forte correlação negativa (r ≈ –0.76)
    - Cooling Load & Relative Compactness: correlação positiva moderada (r ≈ +0.51)
    - Surface Area & Heating Load: correlação negativa moderada (r ≈ –0.43)
    - Glazing Area & Cooling Load: correlação positiva moderada (r ≈ +0.42)

Insights:

- Mais compactos ⇒ menor Heating Load, porém maior Cooling Load.
- Maior área de envidraçamento ⇒ maior demanda de resfriamento, mas impacto menor no aquecimento.

6. Relações Bivariadas

- **Relative Compactness vs. Heating Load**: clara tendência decrescente, sugerindo alta importância dessa variável no modelo.

- **Glazing Area vs. Cooling Load**: dispersão moderada, mas com tendência crescente.

- **Overall Height vs. Heating/Cooling Load**: edifícios de 7 m apresentam ligeiramente cargas mais altas de resfriamento devido ao maior volume interno.

---

7. Outliers e Observações Especiais

- Não há valores extremos extremos (z-score > 3) além de algumas amostras com Cooling Load > 40 kWh/m², que serão mantidas no modelo, pois representam cenários de alto risco térmico.

- Todas as variáveis estão na mesma escala de grandeza (áreas em m², alturas em m, cargas em kWh/m²), facilitando a aplicação direta de algoritmos sem padronização ou com padronização simples (por exemplo, StandardScaler).

---

8. Considerações Finais e Próximos Passos

1. **Pré-processamento**
    - Aplicar StandardScaler ou MinMaxScaler para métodos sensíveis à escala.
    - Dividir dados em treino (70 %) e teste (30 %) garantindo estratificação conforme distribuições de saída.
2. **Seleção de Modelo**
    - Começar com regressão linear regularizada (Lasso, Ridge) para avaliar importância de variáveis.
    - Avançar para modelos não lineares (Random Forest, Gradient Boosting) se necessário.

3. **Validação**

    - Cross-validation k-fold (k=5 ou 10) para estimar desempenho e evitar overfitting.

    - Métricas: RMSE, MAE e R^2 para cada saída (Heating e Cooling Load).