Projeto Modelagem por Regressão Linear Múltipla (MRLM)

Grupo:
- Luiz da Costa Araújo Bronzeado Neto - 123110804
- Leonardo Mota Meira Filho - 123110635

1. Introdução e Objetivo da Análise

Este relatório apresenta os resultados de uma análise de regressão múltipla cujo objetivo é a modelagem estatística do consumo de energia em unidades residenciais. O estudo busca desenvolver um modelo robusto para identificar e quantificar a relação entre o consumo de energia (variável resposta) e um conjunto de variáveis preditoras, que englobam características do imóvel, aspectos demográficos e padrões de uso de equipamentos. A finalidade é obter um modelo com bom poder explicativo e preditivo, cujas inferências sejam estatisticamente válidas e de relevância prática no contexto de gestão energética.

2. Descrição da Base de Dados

A análise parte de uma base de dados com 1997 observações. A estrutura geral dos dados, incluindo as medidas de tendência central e dispersão de cada variável, pode ser observada no sumário inicial.

Saída do Código R: `summary(dados)`

```
 consumo_energia num_moradores      area_m2        temperatura_media
 Min.   :129.2   Min.   :1.000   Min.   :  5.244   Min.   :10.79
 1st Qu.:219.7   1st Qu.:2.000   1st Qu.:100.866   1st Qu.:19.99
 Median :244.5   Median :3.000   Median :120.109   Median :21.90
 Mean   :244.8   Mean   :3.476   Mean   :120.018   Mean   :21.94
 3rd Qu.:271.1   3rd Qu.:5.000   3rd Qu.:139.412   3rd Qu.:23.92
 Max.   :351.6   Max.   :6.000   Max.   :235.553   Max.   :32.64
 NA's   :10
 renda_familiar  uso_ar_condicionado    tipo_construcao equipamentos_eletro
 Min.   :-1242   Não: 784            Apartamento: 586   Min.   : 1.0
 1st Qu.: 4600   Sim:1213            Casa       :1411   1st Qu.: 8.0
 Median : 5920                                         Median :10.0
 Mean   : 5970                                         Mean   :10.1
```

3rd Qu.: 7365 3rd Qu.:12.0

Max. :12168 Max. :22.0

NA's :5

potencia_total_equipamentos

Min. :-1.122

1st Qu.: 9.156

Median :11.961

Mean :12.126

3rd Qu.:14.827

Max. :26.991

A inspeção da integridade dos dados revelou a presença de valores ausentes, concentrados nas variáveis **`consumo_energia`** e **`renda_familiar`**. Não foram encontradas observações duplicadas.

A análise de outliers, realizada através do método do intervalo interquartil (IQR), foi aplicada a todas as variáveis numéricas para identificar observações discrepantes. Os resultados indicaram a presença de outliers em diversas variáveis, com as seguintes contagens: **19** em **`area_m2`**, **17** em **`temperatura_media`**, **14** em **`equipamentos_eletro`**, **10** em **`potencia_total_equipamentos`**, **9** em **`renda_familiar`** e **4** na variável resposta **`consumo_energia`**. A variável **`num_moradores`** foi a única que não apresentou outliers por este critério.

A existência desses pontos é visualmente confirmada pelos boxplots apresentados na análise univariada. De maior relevância, contudo, são os dados claramente inconsistentes que esta análise revelou. Dentre os **10 outliers** de **`potencia_total_equipamentos`** e os **9** de **`renda_familiar`**, foram encontrados registros com valores negativos, o que é economicamente e fisicamente impossível, representando prováveis erros de entrada. Esses erros impactam a qualidade geral do modelo e reforçam a necessidade de uma etapa de depuração dos dados antes da modelagem final.

Saída do Código R: `print(cbind(indice = outliers_renda, ...))` (Exemplo de Inconsistência)

indice valor

[1,] 373 12168.3155

[2,] 1230 -360.2506

[3,] 1247 12076.8866

[4,] 1250 108.3419

[5,] 1264 -347.9823

[6,] 1315 11853.3147

[7,] 1434 -1242.4729

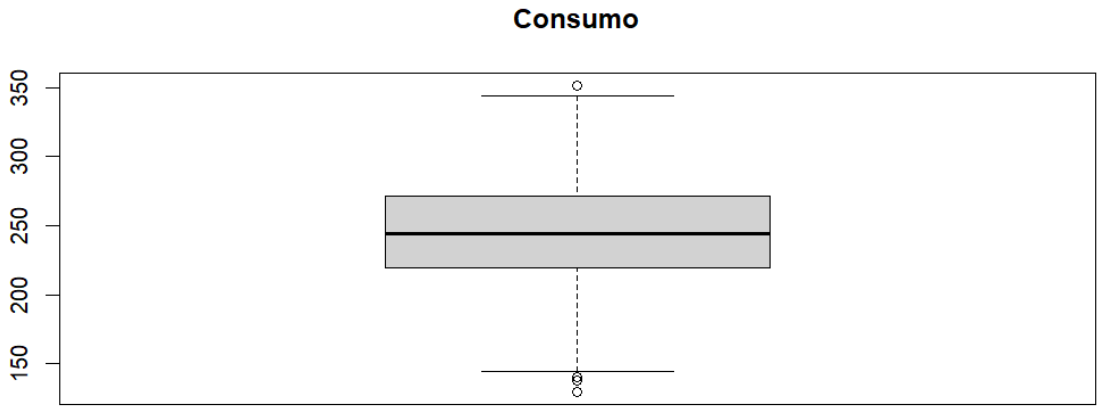
[8,] 1668 -499.5023

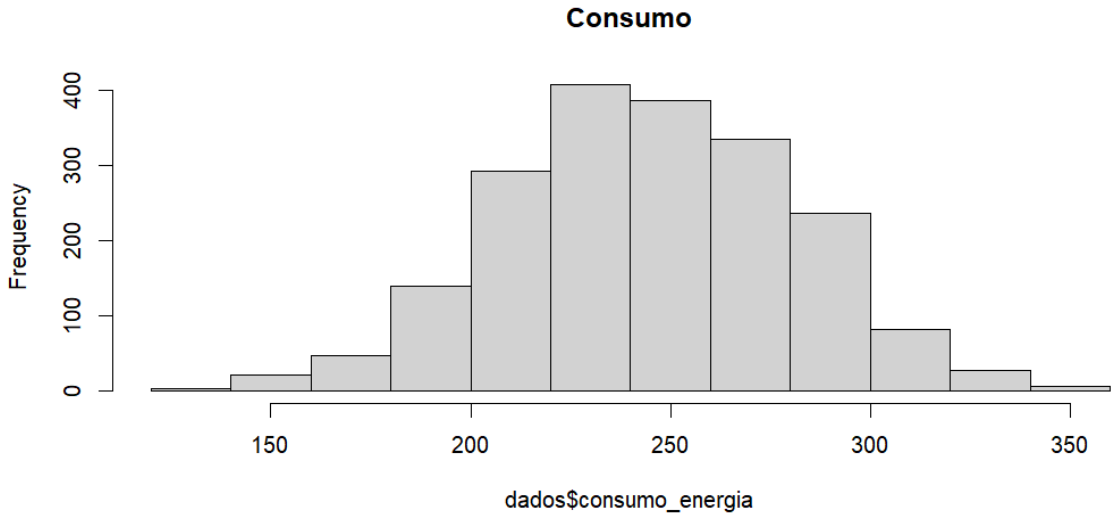
[9,] 1726 -433.1829

3. Análise Exploratória de Dados (EDA) e Análise Univariada

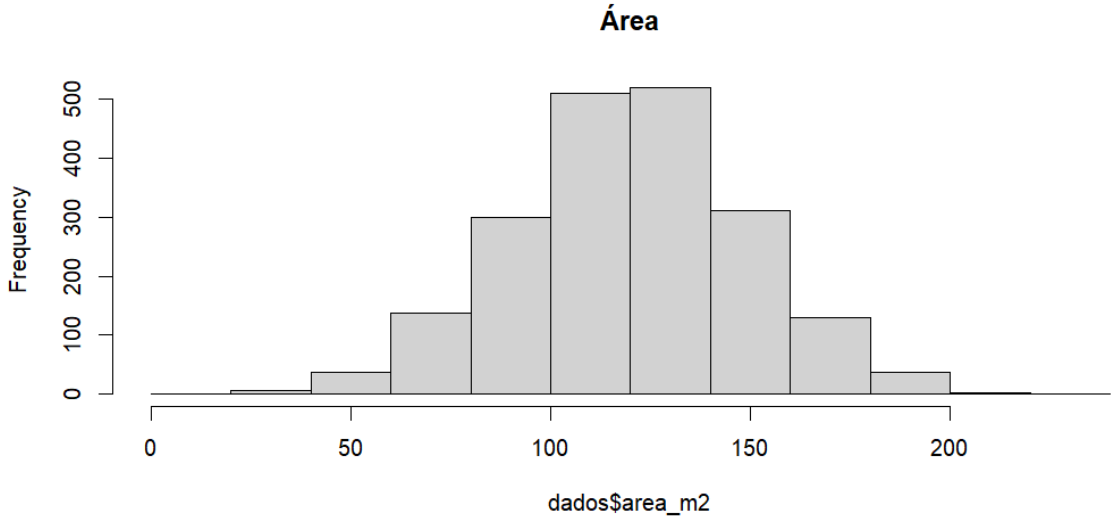
A análise exploratória inicial é um passo fundamental para garantir a qualidade dos dados que fundamentaram o modelo. A primeira etapa consiste em compreender a distribuição de cada variável individualmente (análise univariada).

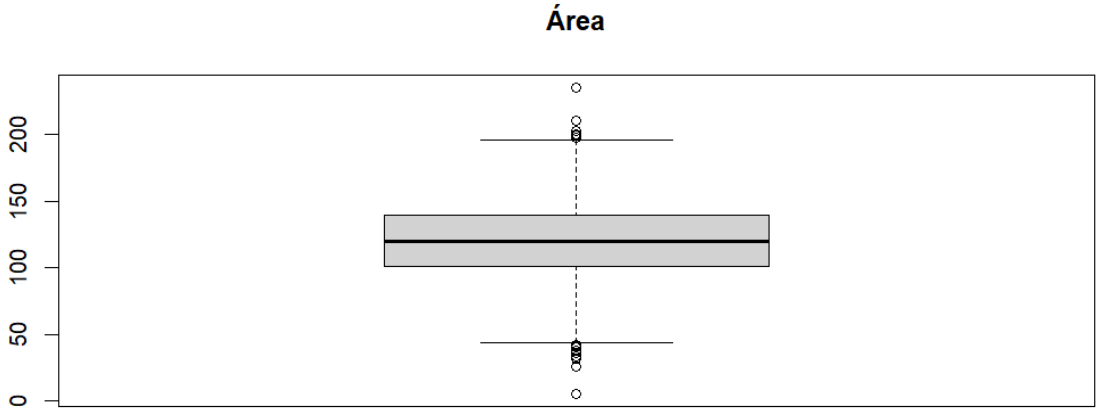
A variável resposta, **`consumo_energia`**, apresenta uma distribuição aproximadamente simétrica e unimodal, com uma concentração de valores em torno de **220-250 kW/h**, como ilustra o histograma. O boxplot correspondente detalha a mediana, os quartis e confirma visualmente a presença de outliers em ambas as caudas da distribuição, corroborando os achados da análise numérica.

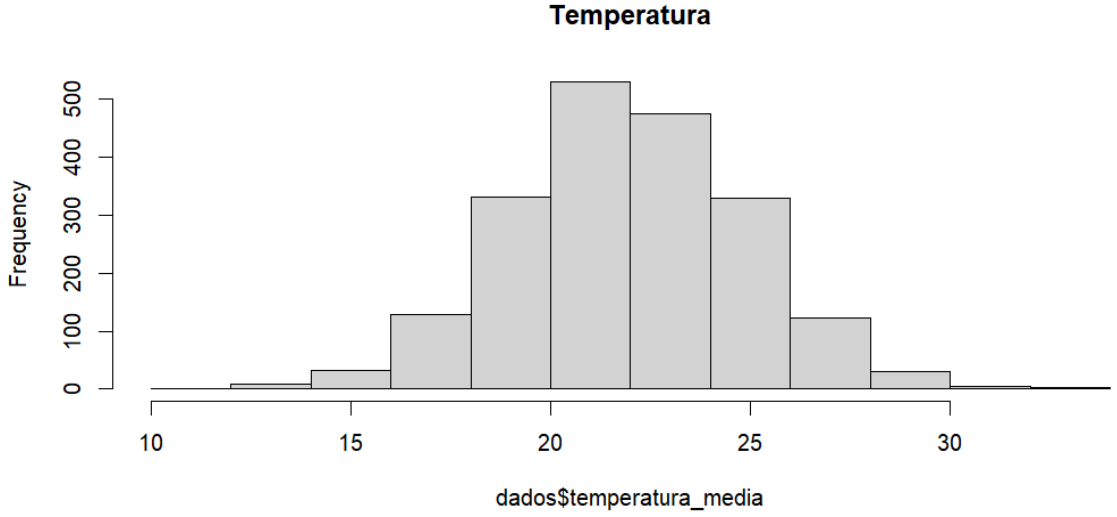


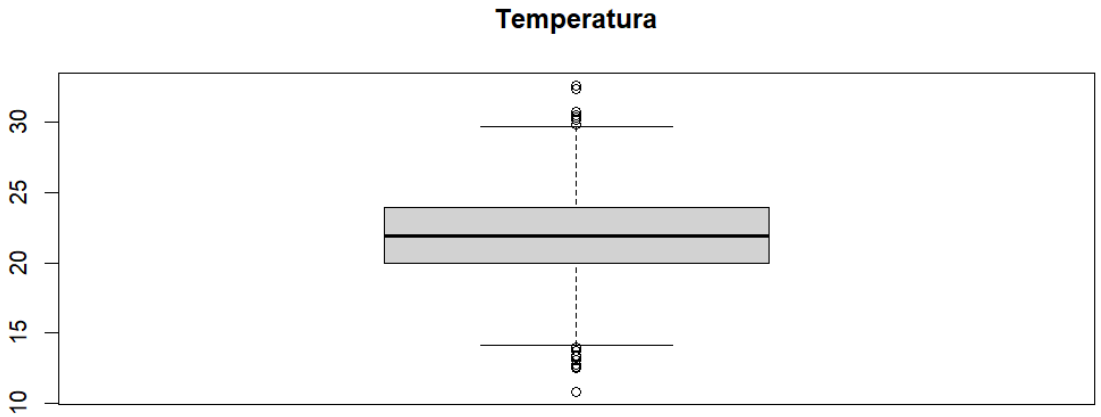


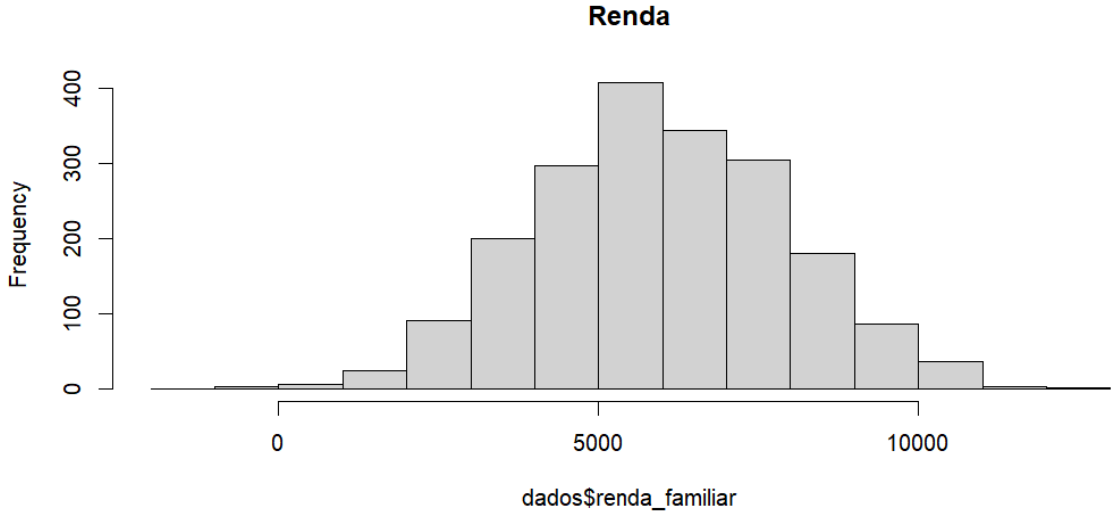
As variáveis preditoras numéricas também foram inspecionadas visualmente. As distribuições de **`area_m2`**, **`temperatura_media`**, **`renda_familiar`**, **`equipamentos_eletro`** e **`potencia_total_equipamentos`** foram analisadas por meio de histogramas para se ter uma noção de sua forma e por boxplots para detalhar suas estatísticas de ordem e identificar outliers. A `Área`, por exemplo, mostra uma distribuição próxima da normalidade, com mediana em torno de 120 m². O boxplot da **Renda** é particularmente informativo, pois, além de mostrar uma distribuição assimétrica à direita, evidencia claramente a presença de outliers na cauda inferior, incluindo os valores negativos que foram apontados como erros de entrada de dados.

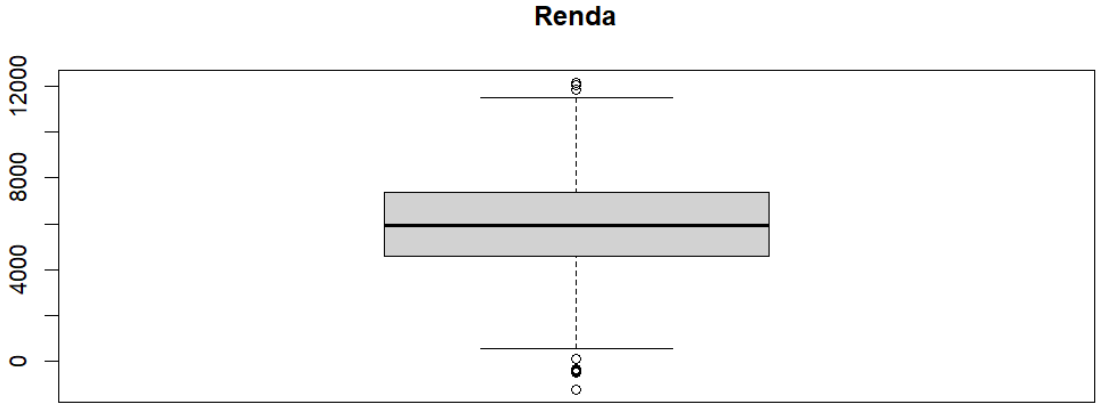

Área
Frequency
0
100
200
300
400
500
0
50
100
150
200
dados$area_m2

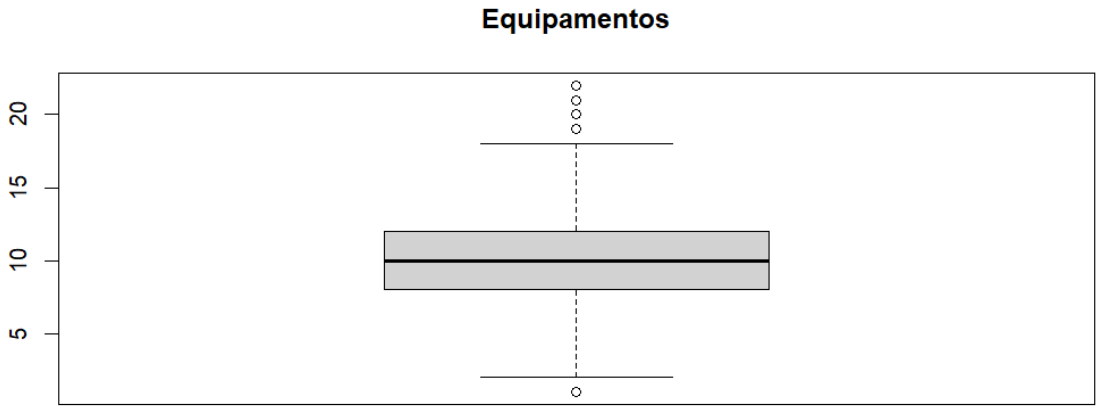

Área
0
50
100
150
200

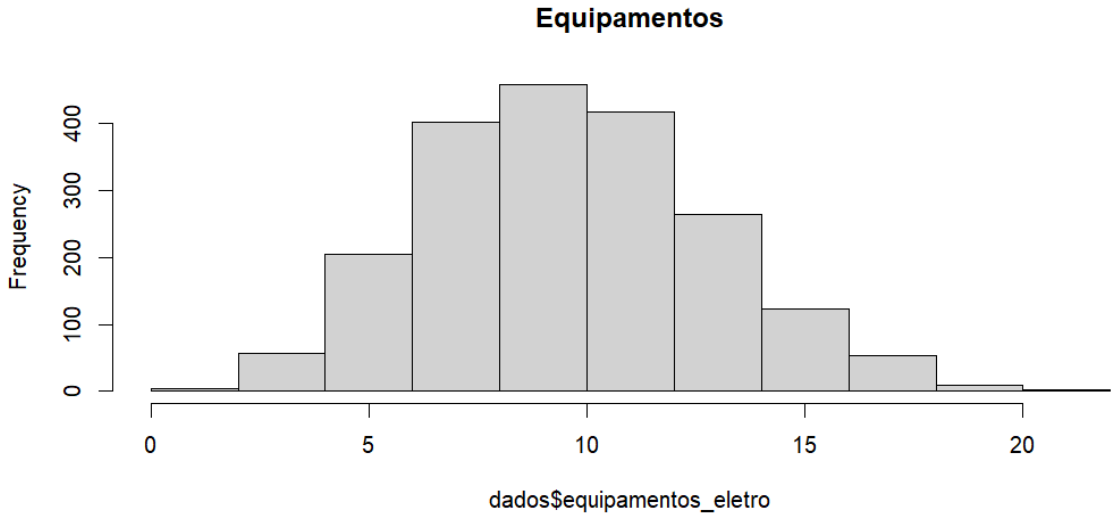

Temperatura
Frequency
0
100
200
300
400
500
10
15
20
25
30
dados$temperatura_media

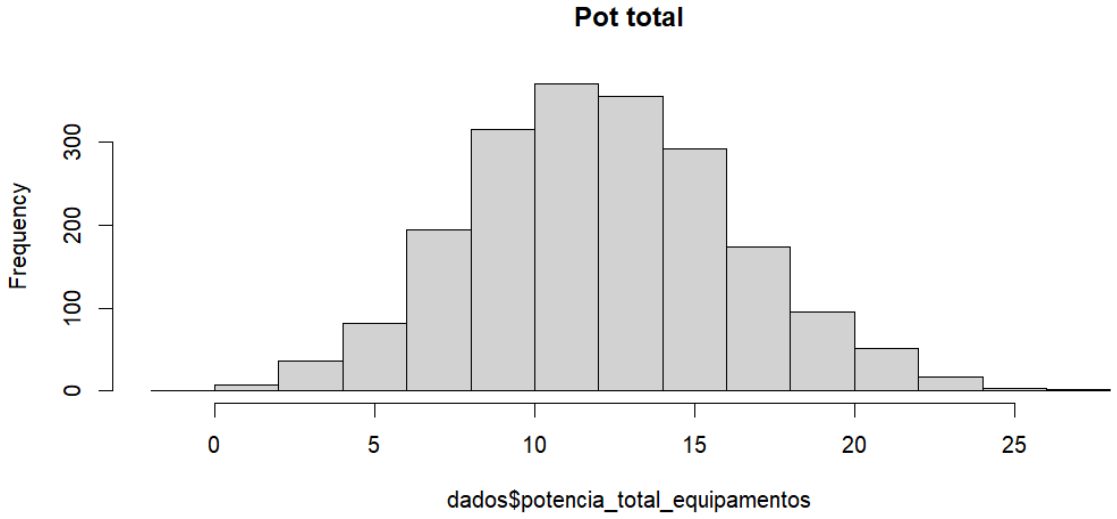

Temperatura
10
15
20
25
30

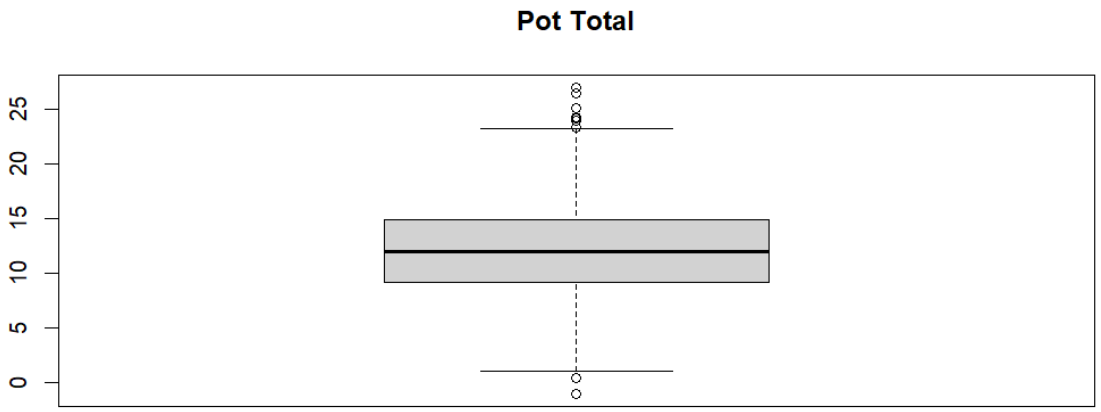

Renda
Frequency
0
100
200
300
400
0
5000
10000
dados$renda_familiar


Renda
0
4000
8000
12000

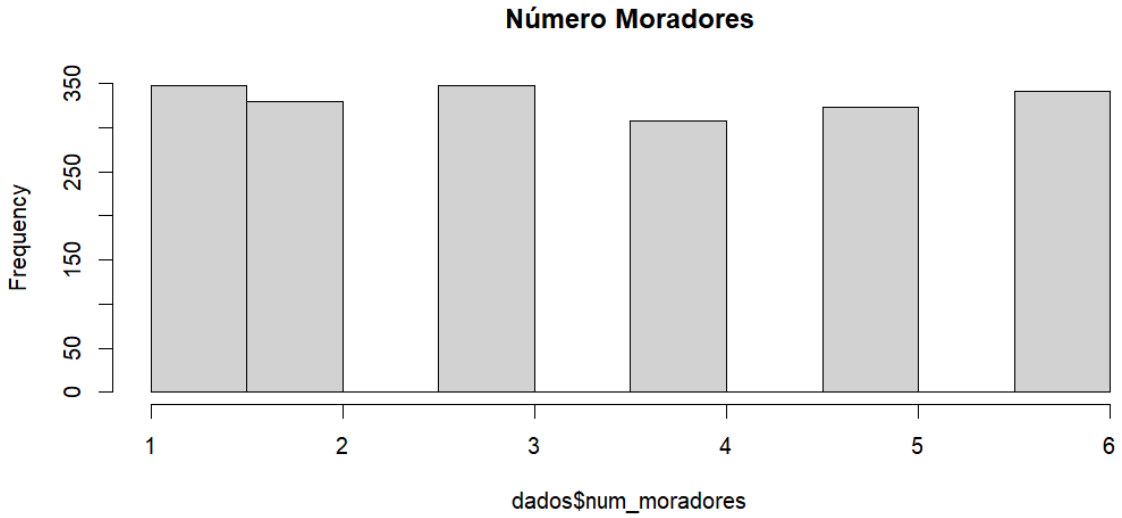

Equipamentos
5
10
15
20

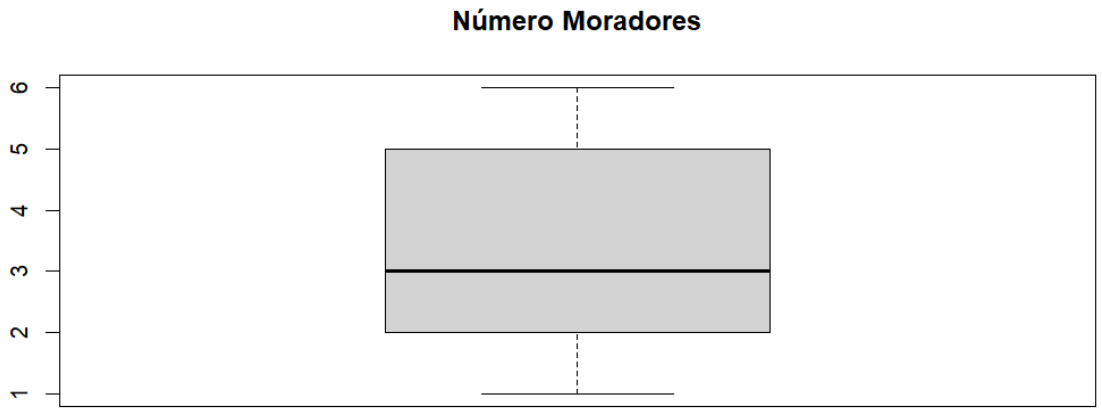

Equipamentos
Frequency
0
100
200
300
400
0
5
10
15
20
dados$equipamentos_eletro

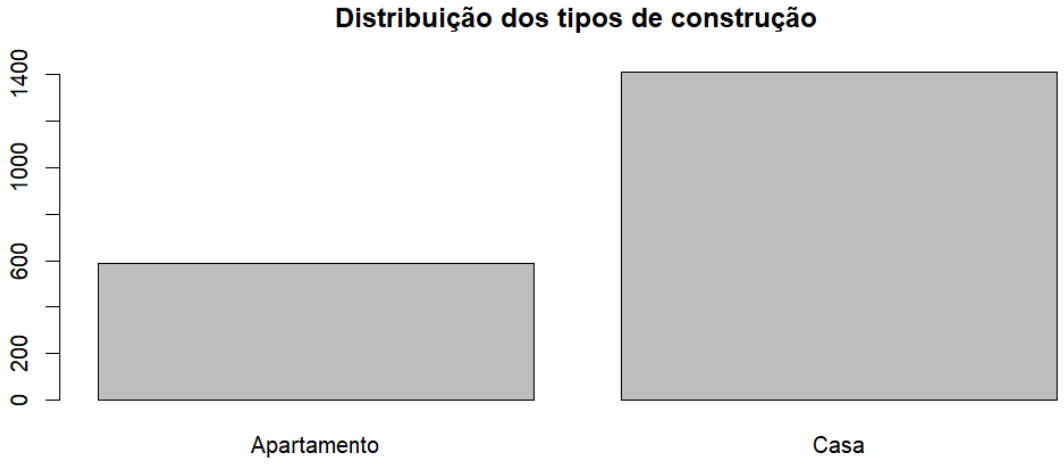

Pot total
Frequency
0
100
200
300
0
5
10
15
20
25
dados$potencia_total_equipamentos

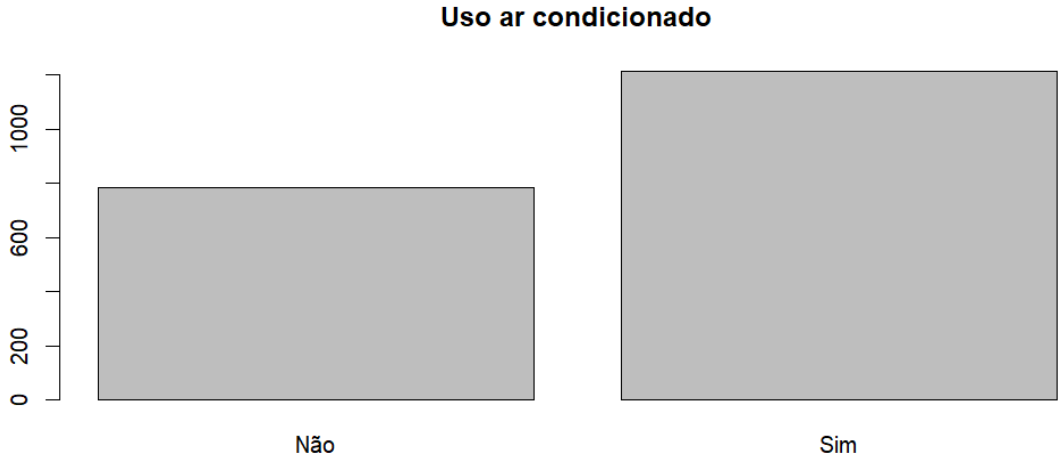


Para as variáveis discretas e categóricas, os gráficos de barras e o boxplot de **`num_moradores`** revelam a composição da amostra. A variável **`num_moradores`** mostra uma distribuição relativamente equilibrada entre as residências de 1 a 6 moradores, com mediana de 3 moradores. A análise de **`uso_ar_condicionado`** revela que há um número maior de domicílios que utilizam o equipamento. Por fim, a amostra é predominantemente composta por residências do tipo "Casa".




Número Moradores
1
2
3
4
5
6

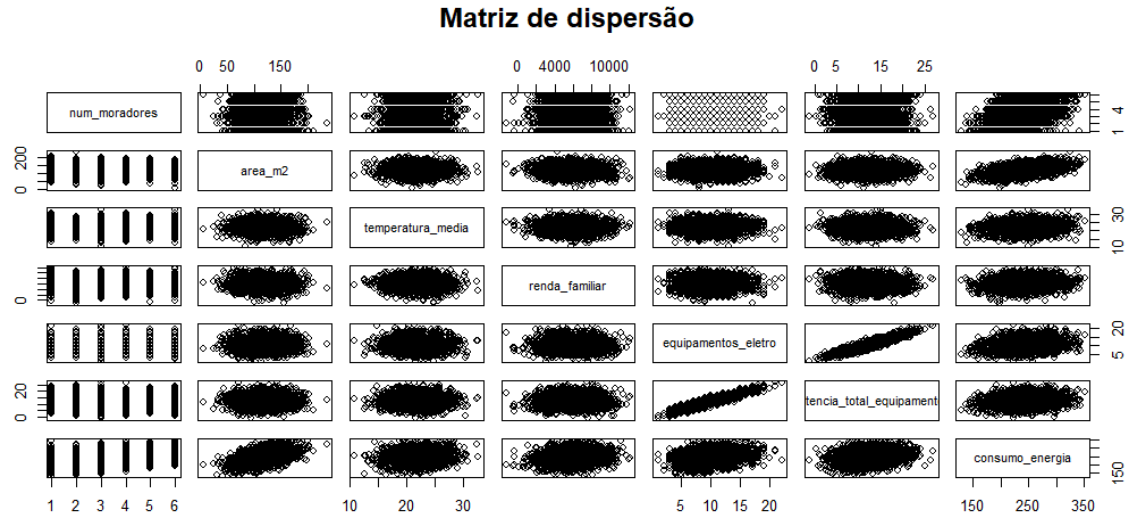

Distribuição dos tipos de construção
0
200
600
1000
1400
Apartamento
Casa


Uso ar condicionado
0
200
600
1000
Não
Sim

Após a análise visual, reitera-se a importância do tratamento dos dados. A investigação identificou valores ausentes e, mais criticamente, valores inconsistentes (negativos) que deveriam ser corrigidos ou removidos. Para este estudo, adotou-se a exclusão automática de casos com dados ausentes, resultando em um tamanho amostral efetivo de **1982 observações**.

4. Análises de Correlação e Multicolinearidade

Para complementar a análise de correlação numérica, a matriz de dispersão a seguir foi gerada, permitindo a visualização simultânea da relação entre todos os pares de variáveis.



A matriz confirma visualmente as associações lineares positivas mais fortes com a variável resposta, notadamente na última coluna, onde se observa a relação entre **`consumo_energia`** e **`area_m2`**. De forma ainda mais contundente, a relação quase perfeitamente linear entre **`equipamentos_eletro`** e **`potencia_total_equipamentos`** (sexta linha, quinta coluna) evidencia a severa multicolinearidade já apontada pela análise de VIF (com valores de 7.78 para ambas as variáveis). Esta redundância de informação inflaciona a variância dos coeficientes estimados, dificultando a interpretação de seus efeitos individuais.

5. Ajuste e Seleção do Modelo

Partindo de um modelo inicial completo, foi aplicado um procedimento de seleção de variáveis do tipo *stepwise* com base no Critério de Informação de Akaike (AIC). O algoritmo convergiu para um modelo que excluiu a variável **`tipo_construcao`**.

O modelo final selecionado apresentou um coeficiente de determinação ajustado (R^2-ajustado) de 0.7001, indicando que aproximadamente 70% da variabilidade amostral do consumo de energia é explicada pelo conjunto de preditores. A significância global do modelo foi atestada pelo alto valor da estatística F (p-valor < 2.2e-16).

6. Análise e Interpretação do Modelo Selecionado

O modelo final ajustado é representado pela seguinte equação:

E(Consumo) = 42.18 + 8.27*(**num_moradores**) + 0.61*(**area_m2**) + 1.66*(**temperatura_media**) + 0.003*(**renda_familiar**) + 31.56*(**uso_ar_condicionadoSim**) + 5.25*(**equipamentos_eletro**) - 2.18*(**potencia_total_equipamentos**)

A interpretação de seus coeficientes deve ser feita com cautela. O coeficiente para **area_m2** (0.61) sugere que, para cada metro quadrado adicional, o consumo médio de energia aumenta em 0.61 kW/h, mantendo as demais variáveis constantes. O coeficiente para **uso_ar_condicionado** == Sim (31.56) indica que residências que utilizam ar condicionado têm um consumo médio 31.56 kW/h maior do que aquelas que não utilizam, ceteris paribus. O sinal negativo para **potencia_total_equipamentos** é um sintoma direto da multicolinearidade, tornando a interpretação isolada deste coeficiente impraticável.

7. Verificação dos Pressupostos do Modelo

A validação do modelo, etapa crucial para assegurar a confiabilidade das conclusões, foi realizada por meio de testes estatísticos formais aplicados aos resíduos. A análise confirmou que os pressupostos fundamentais do Modelo de Regressão Linear Múltipla (MRLM) foram atendidos, conferindo validade às inferências do estudo.

7.1 Normalidade dos Resíduos

Para verificar se os resíduos do modelo seguem uma distribuição normal, foi aplicado o teste de Shapiro-Wilk. A hipótese nula (H_0) deste teste é que os dados são normalmente distribuídos.

Saída do R:

```
Shapiro-Wilk normality test

data:  residuals(modelo_step)
W = 0.99928, p-value = 0.6577
```

Análise: A saída do teste indicou um **p-valor de 0.6577**. Como este valor é significativamente maior que o nível de significância de 0.05, não há evidências para rejeitar a hipótese nula. Portanto, o pressuposto de normalidade dos resíduos é considerado **atendido**.

7.2 Homocedasticidade dos Resíduos

A condição de homocedasticidade (variância constante dos erros) foi avaliada pelo teste de Breusch-Pagan. A hipótese nula (H_0) deste teste é que a variância dos resíduos é constante.

Saída do R:

```
studentized Breusch-Pagan test

data:  modelo_step
BP = 7.0164, df = 7, p-value = 0.4272
```

Análise: O teste resultou em um **p-valor de 0.4272**. Sendo este valor superior a 0.05, falhamos em rejeitar a hipótese nula. Isso indica que não há evidência de heterocedasticidade, confirmando que o pressuposto de variância constante dos erros é **atendido**.

7.3 Independência dos Resíduos

A ausência de autocorrelação entre os resíduos foi verificada pelo teste de Durbin-Watson. A hipótese nula (H_0) é que não há autocorrelação.

Saída do R:

```
Durbin-Watson test

data:  modelo_step
DW = 1.9653, p-value = 0.2196
```

Análise: O **p-valor de 0.2196** está bem acima do limiar de 0.05, levando à não rejeição da hipótese nula. Conclui-se que os resíduos são independentes e não apresentam problemas de autocorrelação.

7.4 Resumo da Validação

- Por fim, a análise de diagnóstico através destes testes estatísticos formais demonstrou que o modelo **`modelo_step`** atende aos pressupostos de normalidade, homocedasticidade e independência dos resíduos, o que confere um alto grau de confiabilidade às inferências, testes de hipótese e intervalos de confiança apresentados neste relatório.

8. Utilização do Modelo para Fins Preditivos

O modelo ajustado pode ser utilizado para estimar o consumo de energia para novas observações. Para ilustrar, consideram-se dois cenários:

- **Cenário 1 (xh=1):** Residência com 2 moradores, 60 m², temperatura média de 25°C, renda de R$ 4000, sem ar condicionado, 8 equipamentos, potência de 10 kW.
- **Cenário 2 (xh=2):** Residência com 4 moradores, 150 m², temperatura média de 28°C, renda de R$ 9000, com ar condicionado, 15 equipamentos, potência de 18 kW.

(a) Previsões e Intervalos de Confiança para o Valor Médio Esperado E(Yh|xh)

Este intervalo estima a **média** de consumo de **todas** as residências com as características de cada cenário. As previsões pontuais para o consumo médio são:

- **Cenário 1:** 169.48 kWh
- **Cenário 2:** 312.31 kWh

Interpretação Prática: Para o Cenário 1, a melhor estimativa para o consumo médio é de 169.48 kWh. Um intervalo de confiança de 95% informaria a faixa de valores plausíveis para esta média populacional.

(b) Previsões e Intervalos de Predição para uma Observação Futura Yh

Este intervalo estima o valor de consumo para uma **única** nova observação, sendo sempre mais largo que o intervalo de confiança. As previsões pontuais são as mesmas do item anterior.

Interpretação Prática: Para o Cenário 2, a melhor estimativa para o consumo de uma residência específica é de 312.31 kWh. O intervalo de predição de 95% forneceria uma gama de valores plausíveis para este caso individual, sendo de maior utilidade para planejamento e previsão em nível micro.

9. Conclusão

Após extensiva análise dos dados, desenvolveu-se um modelo de regressão linear múltipla capaz de explicar aproximadamente 70% da variabilidade no consumo de energia residencial (R^2 ajustado = 0,7001). O modelo indicou de forma clara que a área do imóvel, o número de moradores e a presença de ar-condicionado são fatores determinantes para o aumento do consumo. Além desses, tanto a temperatura média quanto a renda familiar também apresentaram relevância estatística significativa no processo.

No que tange à robustez metodológica do estudo, cabe destacar a adequada validação dos pressupostos do modelo. Os testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson atestaram, respectivamente, a normalidade dos resíduos, a homocedasticidade e a independência dos erros, assegurando qualidade estatística e rigor às inferências realizadas. Dessa maneira, fortaleceu-se não só o poder preditivo, mas também a confiabilidade do modelo para compreensão dos fatores subjacentes ao consumo energético.

É importante salientar, contudo, algumas limitações observadas. Identificou-se acentuada multicolinearidade entre a quantidade de equipamentos e a potência total, o que inviabiliza análises isoladas desses efeitos. Além disso, a presença de dados inconsistentes na base original sinaliza que procedimentos adicionais de depuração poderiam contribuir para elevar a precisão do modelo.

Em resumo, o modelo final apresenta-se como uma ferramenta estatisticamente validada, consistente e com alto potencial prático. Está apto a estimar o consumo de energia em múltiplos contextos e a fornecer subsídios sólidos para análises e tomadas de decisão no setor energético.