blumenau_apartamentos

thiago_guimaraes_sakata

2024-11-26

pacotes

```
library(readxl)
library(tidyverse)
library(gvlma)
library(skimr)
library(car)
library(janitor)
library(formatR)
library(dplyr)
library(lubridate)
library(MASS)
library(caret)
library(summarytools)
```

pesquisa de mercado

```
apartamento <- read_excel("~/IFC/Documentos/16_estagiario/03_blumenau/06_planilha/01_blumenau_2024_apar
```

estrutura

```
str(apartamento)
```

```
## tibble [24 x 21] (S3: tbl_df/tbl/data.frame)
##  $ n                 : chr [1:24] "01" "02" "03" "04" ...
##  $ endereco          : chr [1:24] "R. João Deola, 150 - Progresso" "R. Berta Weise, 303 - Água Verde
##  $ posicao_geografica: logi [1:24] NA NA NA NA NA NA ...
##  $ valor             : num [1:24] 99756 250000 390000 389800 550000 ...
##  $ fonte             : chr [1:24] "MAK IMÓVEIS Creci: 04465-J-SC" "IMÓVEIS PORTAL LTDA-ME Creci: 478
##  $ data              : POSIXct[1:24], format: "2024-10-10" "2024-10-10" ...
##  $ area              : num [1:24] 38 104 72 82 86 54 78 45 70 70 ...
##  $ quarto            : num [1:24] 2 3 2 3 3 2 2 2 2 2 ...
##  $ sanitario         : num [1:24] 1 2 2 2 2 2 2 1 2 2 ...
```

```
##  $ posicao           : num [1:24] 1 1 0 0 1 0 0 0 1 0 ...
##  $ garagem           : num [1:24] 2 1 1 1 2 1 1 1 2 1 ...
##  $ elevador          : num [1:24] 0 0 0 1 1 1 1 1 1 0 ...
##  $ ano               : num [1:24] 2021 1994 2016 1999 2020 ...
##  $ padrao            : num [1:24] 1 1 2 2 2 2 1 1 2 1 ...
##  $ conservacao       : num [1:24] 2 2 2 3 3 3 3 3 3 1 ...
##  $ lazer             : num [1:24] 1 1 1 1 1 1 1 1 1 0 ...
##  $ via_publica       : num [1:24] 1 1 1 1 1 1 1 1 1 1 ...
##  $ fiscal            : num [1:24] 7.83 10.96 14.6 41.32 20.66 ...
##  $ renda             : num [1:24] 1516 2222 1362 2320 2649 ...
##  $ pesquisa          : num [1:24] 2024 2024 2024 2024 2024 ...
##  $ unitario          : num [1:24] 2625 2404 5417 4754 6395 ...
```

Com base na estrutura do *dataframe* `apartamento`, aqui estão algumas observações e conclusões iniciais:

Estrutura e Dados

1. **Dimensão dos Dados**:
   - O *dataframe* possui 24 observações (linhas) e 21 variáveis (colunas). Isso indica que os dados representam 24 apartamentos com suas respectivas características.
2. **Tipos de Dados**:
   - A maioria das variáveis está no formato numérico (`num`), exceto:
     - `n` e `endereco` são *strings* (`chr`), representando o identificador e o endereço do imóvel.
     - `posicao_geografica` é do tipo lógico (`logi`), mas todos os valores estão como `NA`, sugerindo ausência de dados para essa variável.
     - `data` é uma data em formato POSIXct.
   - As variáveis categóricas (`padrao`, `conservacao`, `posicao`, etc.) estão representadas como valores numéricos, que precisam de interpretação de acordo com a descrição fornecida.
3. **Valores Numéricos**:
   - Algumas variáveis numéricas, como `valor` (valor de venda), `area` (área útil), `unitario` (valor por m²), e `fiscal` (valor fiscal), indicam que há dados quantitativos importantes para análises comparativas.
   - As variáveis categóricas com valores numéricos (como `padrao`, `conservacao`, e `via_publica`) precisam ser analisadas de acordo com as definições específicas.

convertendo as variáveis numéricas em categóricas (fatores)

```r
apartamento$posicao <- as.factor(apartamento$posicao)

apartamento$elevador <- as.factor(apartamento$elevador)

apartamento$padrao <- as.factor(apartamento$padrao)

apartamento$conservacao <- as.factor(apartamento$conservacao)

apartamento$lazer <- as.factor(apartamento$lazer)
```

```r
apartamento$via_publica <- as.factor(apartamento$via_publica)

apartamento$pesquisa <- as.factor(apartamento$pesquisa)
```

resumo

```r
dfSummary(apartamento)
```

```
## Data Frame Summary
## apartamento
## Dimensions: 24 x 21
## Duplicates: 0
##
## ------------------------------------------------------------------------------------------------------------------
## No   Variable              Stats / Values                    Freqs (% of Valid)   Graph
## ---- -------------------- --------------------------------- -------------------- ----------------------
## 1    n                    1. 01                              1 ( 4.2%)
##      [character]          2. 02                              1 ( 4.2%)
##                           3. 03                              1 ( 4.2%)
##                           4. 04                              1 ( 4.2%)
##                           5. 05                              1 ( 4.2%)
##                           6. 10                              1 ( 4.2%)
##                           7. 12                              1 ( 4.2%)
##                           8. 13                              1 ( 4.2%)
##                           9. 16                              1 ( 4.2%)
##                           10. 17                             1 ( 4.2%)
##                           [ 14 others ]                     14 (58.3%)           IIIIIIIIIII
##
## 2    endereco             1. R. 25 de Agosto, 425 - It       1 ( 4.2%)
##      [character]          2. R. Berta Weise, 303 - Águ       1 ( 4.2%)
##                           3. R. Bertha Muller, 95 - Sa       1 ( 4.2%)
##                           4. R. Coruripe, 86 - Água Ve       1 ( 4.2%)
##                           5. R. João Deola, 150 - Prog       1 ( 4.2%)
##                           6. R. João Pessoa, 2615 - Ve       1 ( 4.2%)
##                           7. Rua Amazonas, 1400 - Garc       1 ( 4.2%)
##                           8. Rua Antônio Treis, 988 -        1 ( 4.2%)
##                           9. Rua Farmacêutico Adolfo O       1 ( 4.2%)
##                           10. Rua Frieda Jensen, 1 - It      1 ( 4.2%)
##                           [ 14 others ]                     14 (58.3%)           IIIIIIIIIII
##
## 3    posicao_geografica   All NA's
##      [logical]
##
## 4    valor                Mean (sd) : 500398.2 (256290.2)   21 distinct values     : :
##      [numeric]            min < med < max:                                         : :
##                           99756 < 429500 < 1300000                                 : :
##                           IQR (CV) : 216250 (0.5)                                  : :
##                                                                                   . : : : :   .
##
## 5    fonte                1. Zelt Imóveis LTDA Creci:       6 (25.0%)            IIIII
```

```
##      [character]          2. RD Imóveis Creci: 6559-J-      3 (12.5%)           II
##                           3. Momento Certo Imobiliária      2 ( 8.3%)           I
##                           4. Assur Fernandes II             1 ( 4.2%)
##                           5. Barbieri Negócios Imobili      1 ( 4.2%)
##                           6. Efetiva Consultoria Imobi      1 ( 4.2%)
##                           7. Giancarlo A. Geremias          1 ( 4.2%)
##                           8. Gido Portella Creci: 1824      1 ( 4.2%)
##                           9. IMÓVEIS PORTAL LTDA-ME Cr      1 ( 4.2%)
##                           10. La Vita Imóveis Creci: 04     1 ( 4.2%)
##                           [ 6 others ]                      6 (25.0%)           IIIII
##
## 6    data                 1. 2024-10-10                     24 (100.0%)         IIIIIIIIIIIIIIIIIIII
##      [POSIXct, POSIXt]
##
## 7    area                 Mean (sd) : 78.4 (26.2)           21 distinct values      :
##      [numeric]            min < med < max:                                          :
##                           38 < 73.5 < 167                                           :
##                           IQR (CV) : 20.5 (0.3)                                     : :
##                                                                                 . : : : . .   .
##
## 8    quarto               Min  : 2                          2 : 17 (70.8%)      IIIIIIIIIIIIII
##      [numeric]            Mean : 2.3                        3 :  7 (29.2%)      IIIII
##                           Max  : 3
##
## 9    sanitario            Mean (sd) : 2.1 (0.8)             1 :  4 (16.7%)      III
##      [numeric]            min < med < max:                  2 : 15 (62.5%)      IIIIIIIIIIII
##                           1 < 2 < 4                         3 :  3 (12.5%)      II
##                           IQR (CV) : 0 (0.4)                4 :  2 ( 8.3%)      I
##
## 10   posicao              1. 0                              15 (62.5%)          IIIIIIIIIIII
##      [factor]             2. 1                               9 (37.5%)          IIIIIII
##
## 11   garagem              Min  : 1                          1 : 13 (54.2%)      IIIIIIIIII
##      [numeric]            Mean : 1.5                        2 : 11 (45.8%)      IIIIIIIII
##                           Max  : 2
##
## 12   elevador             1. 0                               5 (20.8%)          IIII
##      [factor]             2. 1                              19 (79.2%)          IIIIIIIIIIIIIII
##
## 13   ano                  Mean (sd) : 2014 (11.8)           14 distinct values          :
##      [numeric]            min < med < max:                                              :
##                           1970 < 2017 < 2024                                            : .
##                           IQR (CV) : 6.5 (0)                                            : :
##                                                                                   .   : . : :
##
## 14   padrao               1. 1                               5 (20.8%)          IIII
##      [factor]             2. 2                              17 (70.8%)          IIIIIIIIIIIII
##                           3. 3                               2 ( 8.3%)          I
##
## 15   conservacao          1. 1                               1 ( 4.2%)
##      [factor]             2. 2                               5 (20.8%)          IIII
##                           3. 3                              18 (75.0%)          IIIIIIIIIIIIII
##
## 16   lazer                1. 0                               4 (16.7%)          III
```

```
##      [factor]            2. 1                            20 (83.3%)          IIIIIIIIIIIIIIII
##
## 17   via_publica         1. 1                            24 (100.0%)         IIIIIIIIIIIIIIIIIII
##      [factor]
##
## 18   fiscal              Mean (sd) : 32.8 (25.7)         14 distinct values   :
##      [numeric]           min < med < max:                                    :
##                          7.8 < 21.3 < 103.3                                  :   :
##                          IQR (CV) : 26.3 (0.8)                               : : :
##                                                                              : : :     :
##
## 19   renda               Mean (sd) : 2269.2 (924.9)      23 distinct values        :
##      [numeric]           min < med < max:                                    .     :
##                          880.6 < 2169.6 < 4622                               : : :
##                          IQR (CV) : 1203.5 (0.4)                             : : : :
##                                                                              . : : : : . . . .
##
## 20   pesquisa            1. 2024                         24 (100.0%)         IIIIIIIIIIIIIIIIII
##      [factor]
##
## 21   unitario            Mean (sd) : 6213.1 (1693.7)     24 distinct values        :
##      [numeric]           min < med < max:                                          : .
##                          2403.8 < 6188.7 < 9111.1                                  : : :
##                          IQR (CV) : 2118.9 (0.3)                             .     : : : : .
##                                                                              :     : : : : : .
## ----------------------------------------------------------------------------------------------------
```

A análise do resumo do *dataframe* `apartamento` com a função `dfSummary` permite tirar algumas conclusões sobre a estrutura, a distribuição e a qualidade dos dados. Aqui estão as principais observações:

---

1. Dados ausentes

- A variável `posicao_geografica` está completamente ausente (`NA` em 100% das linhas). Isso indica que essa variável pode não ter sido coletada ou precisa ser preenchida (via geocodificação, por exemplo) para ser utilizada.
- As outras variáveis estão completas, sem valores ausentes, o que é positivo para análises subsequentes.

---

2. Distribuição das variáveis

Variáveis categóricas (fatores):

- **`posicao`**: A maioria dos apartamentos não tem sacada voltada para a rua (62,5% com valor `0`).
- **`elevador`**: A grande maioria dos prédios tem elevador (79,2%), o que pode ser um indicador de padrão mais elevado.
- **`padrao`**: A maioria dos apartamentos tem padrão construtivo "normal" (70,8% com valor `2`).
- **`conservacao`**: O estado de conservação predominante é "excelente" (75,0% com valor `3`).
- **`lazer`**: A maioria dos prédios oferece área de lazer (83,3% com valor `1`).
- **`via_publica`**: Todos os apartamentos estão em vias pavimentadas (100% com valor `1`).

Variáveis numéricas:

- **`valor` (valor de venda):**
  - Média: R$ 500.398,20
  - Amplitude: de R$ 99.756,00 a R$ 1.300.000,00
  - A mediana (R$ 429.500,00) é menor que a média, indicando uma distribuição levemente assimétrica à direita.
- **`area` (área útil):**
  - Média: 78,4 m²
  - Amplitude: de 38 m² a 167 m²
  - A maioria dos apartamentos tem tamanho mediano (73,5 m²).
- **`unitario` (valor por m²):**
  - Média: R$ 6.213,10/m²
  - Amplitude: de R$ 2.403,80/m² a R$ 9.111,10/m²
  - Valores com alta dispersão (desvio padrão de R$ 1.693,70).
- **`idade` (ano de construção):**
  - Média: 2014
  - Amplitude: de 1970 a 2024
  - A maioria dos edifícios é relativamente nova, com mediana em 2017.
- **`fiscal` (valor fiscal):**
  - Média: R$ 32,80/m²
  - Alta variabilidade: de R$ 7,83 a R$ 103,30/m²
  - Discrepância significativa entre os valores fiscais por m².

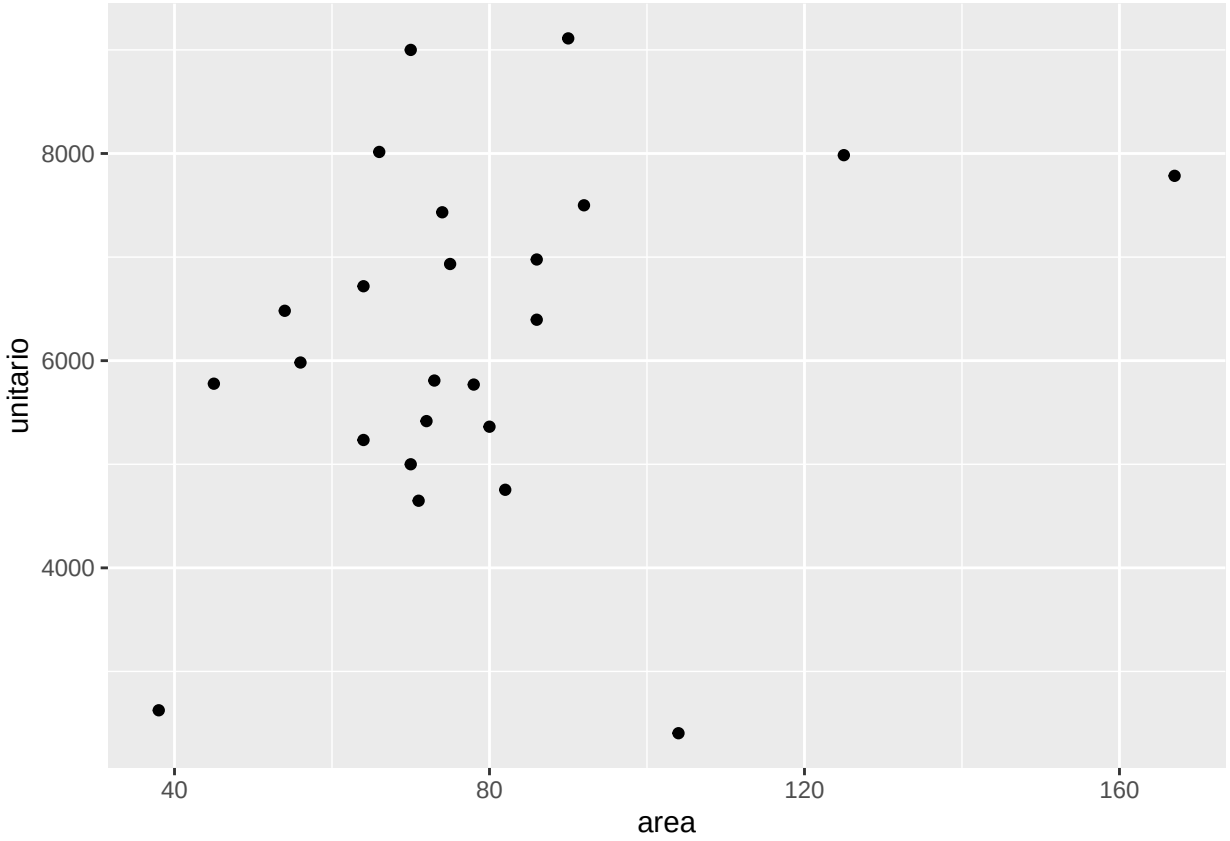
---

3. Observações sobre os dados

1. **Representatividade da amostra**:
   - Há uma boa variedade de apartamentos em termos de valores e características, com 24 observações e diferentes atributos.
   - A distribuição de valores como `valor`, `unitario` e `area` é ampla, o que é bom para análises estatísticas.
2. **Condições do mercado imobiliário**:
   - A relação entre o preço por m² (`unitario`) e outros atributos, como padrão (`padrao`), conservação (`conservacao`), e presença de lazer (`lazer`), será importante para avaliar como essas variáveis impactam os preços.
   - A existência de elevador, padrão construtivo mais elevado e estado de conservação "excelente" são características predominantes, indicando uma tendência de apartamentos com melhores condições.
3. **Dados fiscais**:
   - A discrepância entre o valor fiscal e o valor comercial pode ser explorada em análises de avaliação de imóveis para verificar o impacto das condições de mercado em relação aos parâmetros oficiais.

area x unitario

```r
ggplot(data = apartamento) +
  geom_point(mapping = aes(x = area, y = unitario))
```



Com base no gráfico de dispersão (*scatter plot*) que relaciona a variável **`area`** (área útil) no eixo X e **`unitario`** (valor por m²) no eixo Y, podemos observar o seguinte:

1. Tendência Geral

- Não há uma tendência clara de relação linear entre **`area`** e **`unitario`**.
- Os valores de **`unitario`** (preço por m²) parecem apresentar maior variabilidade em áreas menores, enquanto os valores se tornam mais concentrados em áreas maiores.

---

2. Variabilidade

- **Apartamentos menores** (entre 40 e 80 m²):
  - Os valores de **`unitario`** variam amplamente, de cerca de R$ 2.500/m² a R$ 9.000/m².
- **Apartamentos maiores** (acima de 100 m²):
  - Os valores de **`unitario`** tendem a se estabilizar entre R$ 6.000/m² e R$ 8.000/m².

---

3. Possíveis Interpretações

- Apartamentos menores podem ser mais sensíveis a outros fatores, como localização, padrão construtivo, ou estado de conservação, o que pode justificar a maior variação no preço por m².
- Apartamentos maiores têm uma faixa de preço por m² mais consistente, possivelmente devido à menor demanda relativa ou menor variação de características qualitativas.

transformando area = ln(area)

```
apartamento$ln_area <- log(apartamento$area)
```

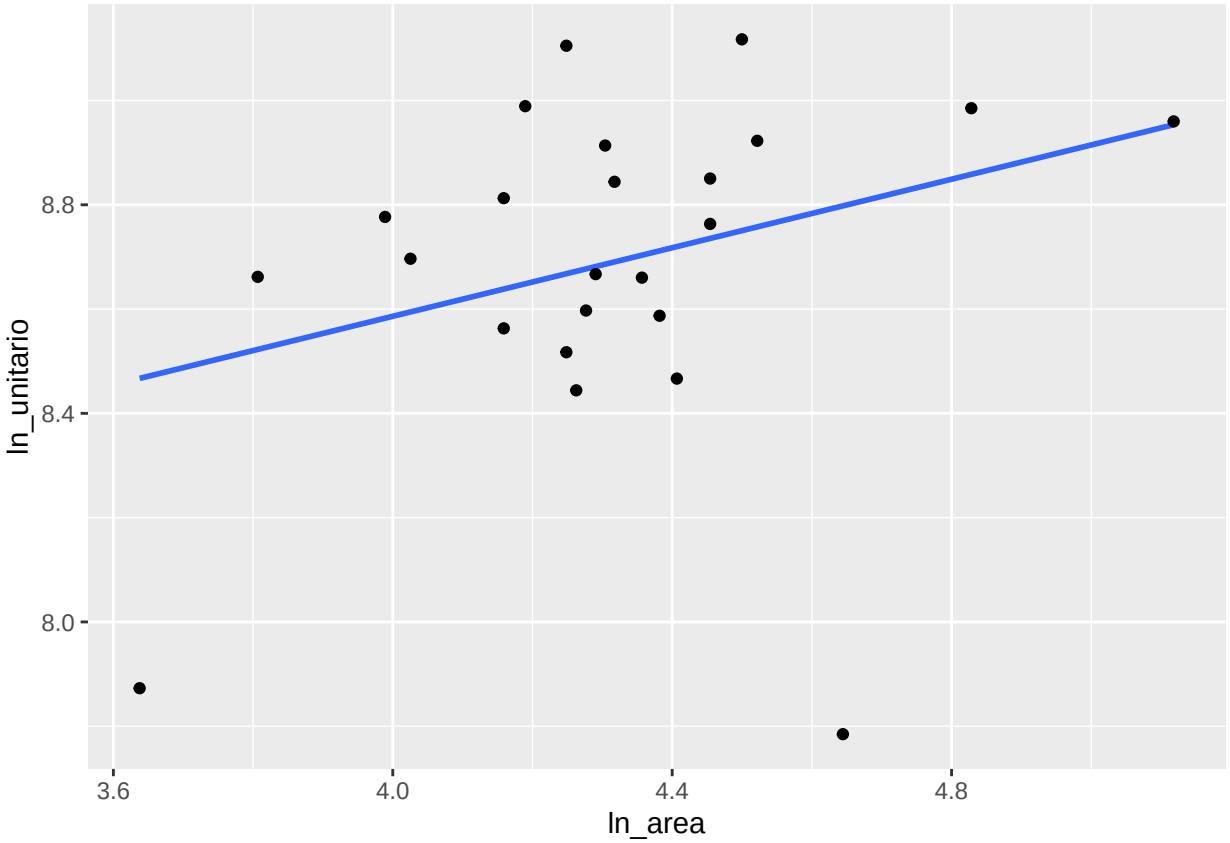
transformando unitario = ln(unitario)

```
apartamento$ln_unitario <- log(apartamento$unitario)
```

ln_area x ln_unitario

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = ln_area, y = ln_unitario), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = ln_area, y = ln_unitario))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico de dispersão e na linha de tendência linear ajustada (log-log plot) que relaciona **ln_area** (log da área útil) no eixo X com **ln_unitario** (log do valor por m²) no eixo Y, podemos tirar as seguintes conclusões:

---

1. Tendência Positiva

- A linha de regressão indica uma relação **positiva** entre **ln_area** e **ln_unitario**:
  - Conforme o logaritmo da área aumenta, o logaritmo do valor por m² também tende a aumentar.
  - Em termos reais, apartamentos maiores (em área útil) tendem a ter valores por m² mais altos em uma escala logarítmica, embora a relação seja moderada.

---

2. Redução na Dispersão

- O uso de logaritmos suavizou a dispersão dos dados em comparação ao gráfico original com os valores brutos.
- Isso sugere que uma transformação logarítmica é apropriada para modelar a relação entre **area** e **unitario**, reduzindo possíveis efeitos de heterocedasticidade (variância não constante dos resíduos).

---

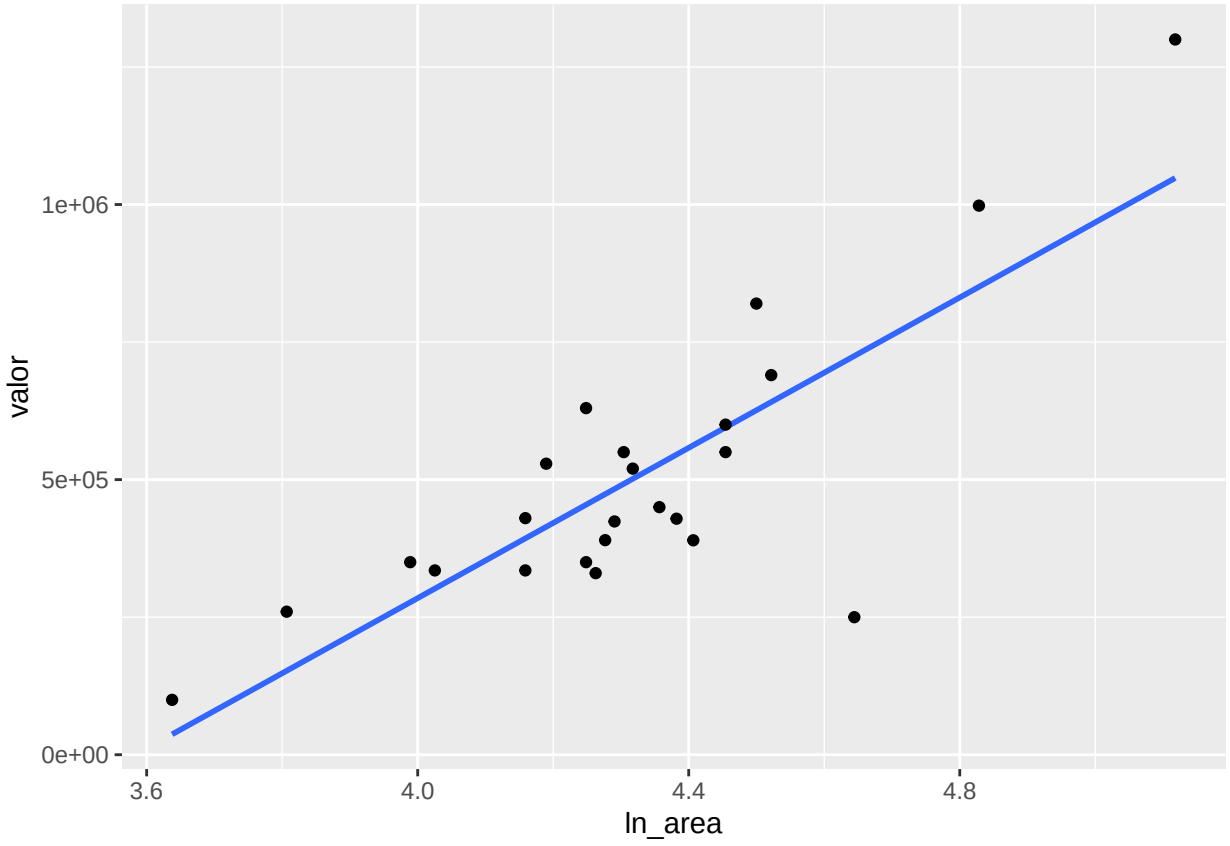
3. Adequação do Modelo

- A linha de tendência ajustada pelo método de regressão linear indica que o modelo log-log é razoável para explicar a relação entre as variáveis.
- Apesar disso, há certa dispersão dos pontos ao redor da linha, o que sugere que a **área** (mesmo em escala logarítmica) não explica completamente as variações no valor por m². Outros fatores podem estar influenciando, como:
  - **Localização** (bairros diferentes).
  - **Padrão construtivo** e **estado de conservação**.
  - **Características adicionais** (elevador, área de lazer, etc.).

ln_area x valor

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = ln_area, y = valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = ln_area, y = valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico que relaciona **ln_area** (logaritmo da área útil) no eixo X e **valor** (valor de venda) no eixo Y, com uma linha de tendência ajustada por regressão linear, podemos tirar as seguintes conclusões:

1. **Relação positiva entre ln_area e valor**

- Há uma relação claramente positiva entre **ln_area** e **valor**:
  - À medida que o logaritmo da área útil aumenta, o valor de venda também aumenta.
  - Isso indica que imóveis maiores tendem a ter valores de venda mais altos, como esperado.

---

2. **Ajuste linear é razoável**

- A linha de tendência ajustada sugere que o modelo linear é razoavelmente adequado para representar essa relação.
- A dispersão dos pontos em torno da linha de regressão é relativamente moderada, mas não perfeita. Isso indica que a **ln_area** é um bom preditor do **valor**, mas outros fatores também podem influenciar o preço de venda.

---

3. **Variabilidade nos valores de venda**

- Para valores intermediários de **ln_area** (aproximadamente entre 4 e 4.4), há uma dispersão maior no **valor**:
  - Isso pode refletir a influência de outros atributos dos imóveis (como localização, padrão construtivo, conservação, etc.) que não estão diretamente relacionados ao tamanho.
- Nos extremos (tanto áreas pequenas quanto grandes), os valores estão mais alinhados com a linha de regressão.

---

4. **Possíveis interpretações**

- Apartamentos maiores, como esperado, apresentam valores de venda mais altos, mas não de forma diretamente proporcional. A relação logarítmica sugere que o impacto da área no preço de venda diminui para áreas muito grandes.
- A dispersão indica que **ln_area** não é o único fator determinante para o valor de venda; outros fatores (qualitativos e quantitativos) precisam ser incluídos para um modelo mais robusto.

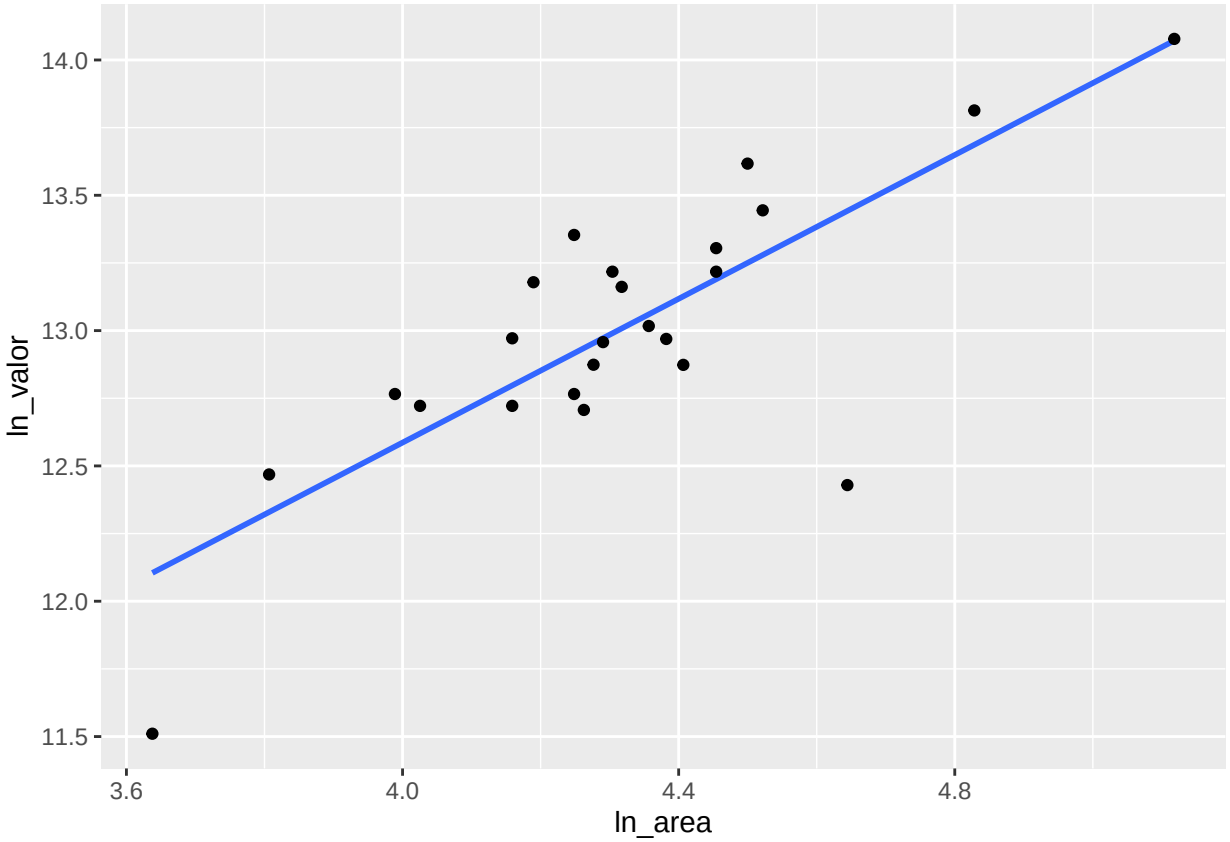
transformando valor = ln(valor)

```
apartamento$ln_valor <- log(apartamento$valor)
```

ln_area x ln_valor

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = ln_area, y = ln_valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = ln_area, y = ln_valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico que relaciona **ln_area** (logaritmo da área útil) no eixo X e **ln_valor** (logaritmo do valor de venda) no eixo Y, com uma linha de regressão ajustada, podemos tirar as seguintes conclusões:

---

1. Relação linear positiva

- Existe uma **forte relação linear positiva** entre **ln_area** e **ln_valor**:
  - Quando o logaritmo da área útil aumenta, o logaritmo do valor de venda também aumenta.
  - Isso confirma que o valor de venda dos apartamentos está relacionado ao tamanho em uma escala logarítmica.

---

2. Melhor ajuste

- O uso de logaritmos para ambas as variáveis reduziu a dispersão e tornou a relação mais linear em comparação com gráficos anteriores.
- A linha de regressão ajustada está bem próxima dos pontos, sugerindo que o modelo linear log-log é adequado para capturar essa relação.

---

3. Variabilidade reduzida

- A dispersão ao longo da linha de regressão é menor, especialmente para valores intermediários de **ln_area** (aproximadamente entre 4 e 4.5), o que reforça a validade da transformação logarítmica para modelar a relação.

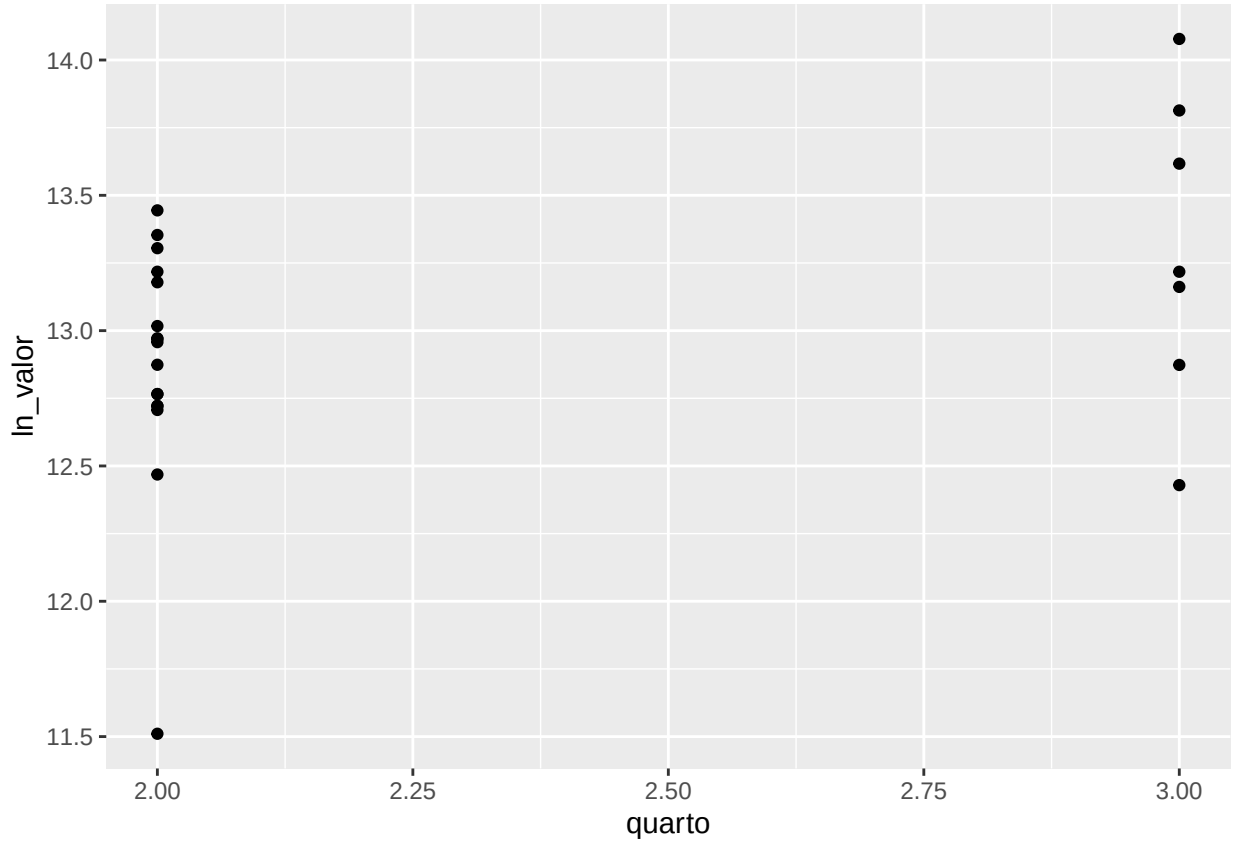
---

4. Interpretação econômica

- Em uma escala log-log, a inclinação da linha de regressão representa a elasticidade do valor de venda em relação à área útil:
  - Um aumento percentual na área útil está associado a um aumento percentual proporcional no valor de venda.
  - Esse comportamento é típico do mercado imobiliário, onde o tamanho do imóvel tem um impacto significativo no preço.

quarto x ln_valor

```
ggplot(data = apartamento) +
  geom_point(mapping = aes(x = quarto, y = ln_valor))
```



Com base no gráfico que relaciona **quarto** (número de quartos) no eixo X e **ln_valor** (logaritmo do valor de venda) no eixo Y, podemos tirar as seguintes conclusões:

---

1. Relação entre número de quartos e valor de venda

- Há uma separação clara nos dados em dois grupos principais: apartamentos com **2 quartos** e **3 quartos**, sendo os de **2 quartos** mais frequentes.
- Não há uma variação significativa de **ln_valor** dentro de cada grupo, indicando que o número de quartos, por si só, não explica totalmente a variabilidade do valor do imóvel.

---

2. Possíveis padrões

- **Apartamentos com 3 quartos** tendem a ter valores mais altos de **ln_valor** em comparação com os de 2 quartos, o que é esperado, já que o número de quartos é geralmente associado a um maior valor de mercado.
- No entanto, a dispersão de **ln_valor** dentro do grupo de 2 quartos é maior, o que sugere que outros fatores, como localização, área útil, padrão construtivo ou estado de conservação, têm maior impacto no valor de venda do que apenas o número de quartos.

---

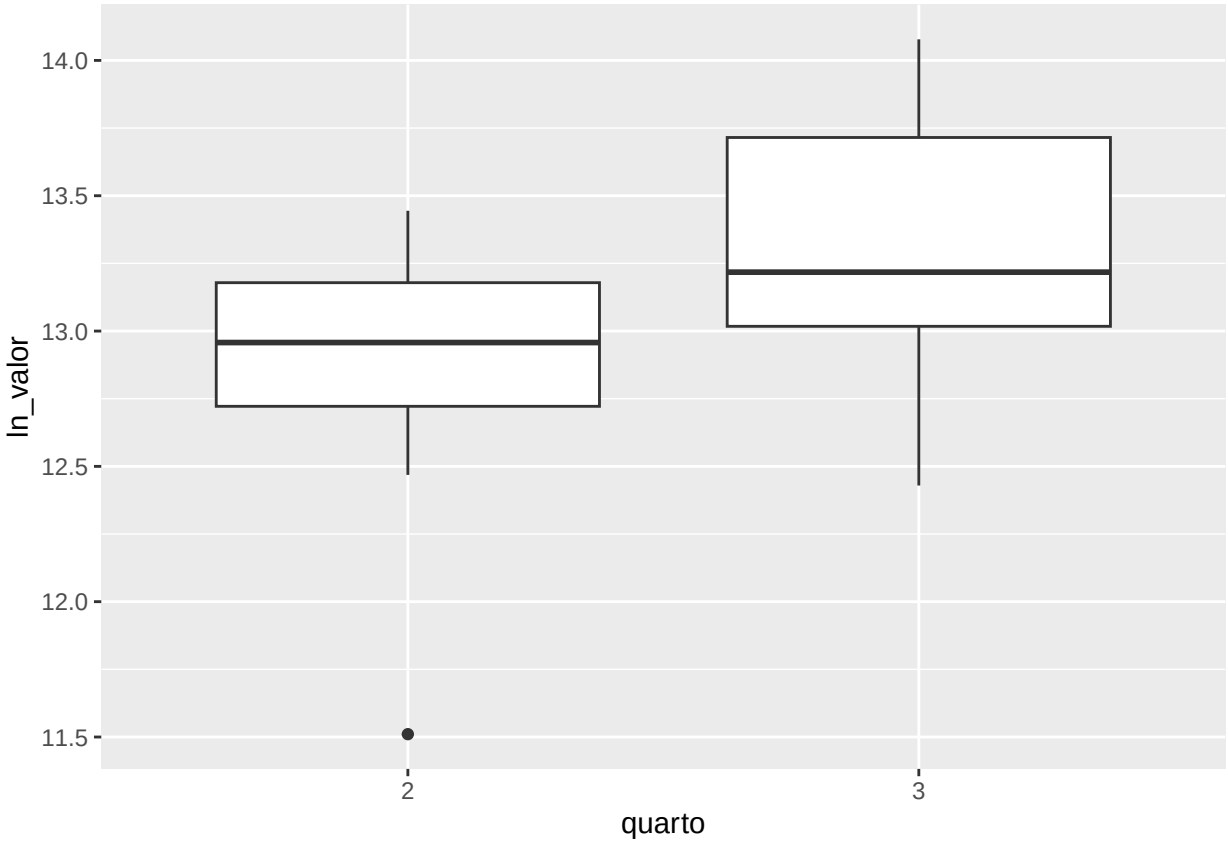
3. Interpretação econômica

- O número de quartos parece ser um fator qualitativo que influencia o valor do imóvel, mas é insuficiente como preditor único, já que os valores de **ln_valor** sobrepõem-se entre os dois grupos.

transformando quarto em variavel categorica

```
apartamento$quarto <- as.factor(apartamento$quarto)

ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = quarto, y = ln_valor))
```



Com base no gráfico de boxplot que relaciona o número de quartos (**quarto**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Diferença entre grupos

- **Apartamentos com 3 quartos** possuem valores médios de **ln_valor** maiores do que os apartamentos com 2 quartos. Isso confirma que o número de quartos está positivamente associado ao valor de venda, como esperado.
- A mediana de **ln_valor** para apartamentos de 3 quartos é mais alta que a dos de 2 quartos, indicando que, em geral, imóveis com mais quartos são mais caros.

---

2. Dispersão e variabilidade

- **Apartamentos com 3 quartos** apresentam maior variabilidade no logaritmo do valor de venda (altura do boxplot maior):
  - Isso pode refletir diferenças em outros fatores, como localização, padrão construtivo ou área útil, que afetam o valor de venda nesses imóveis.
- **Apartamentos com 2 quartos** têm uma dispersão menor em **ln_valor**, mas há um outlier (valor muito baixo), que pode indicar um imóvel com características muito distintas ou uma inconsistência nos dados.

---

3. Sobreposição entre grupos

- Apesar da diferença nas medianas, há uma sobreposição considerável nos valores de **ln_valor** entre os dois grupos:
  - Alguns apartamentos com 2 quartos possuem valores similares ou superiores aos de 3 quartos.
  - Isso reforça a ideia de que o número de quartos não é o único fator determinante do valor de venda e que outras variáveis precisam ser consideradas.

---

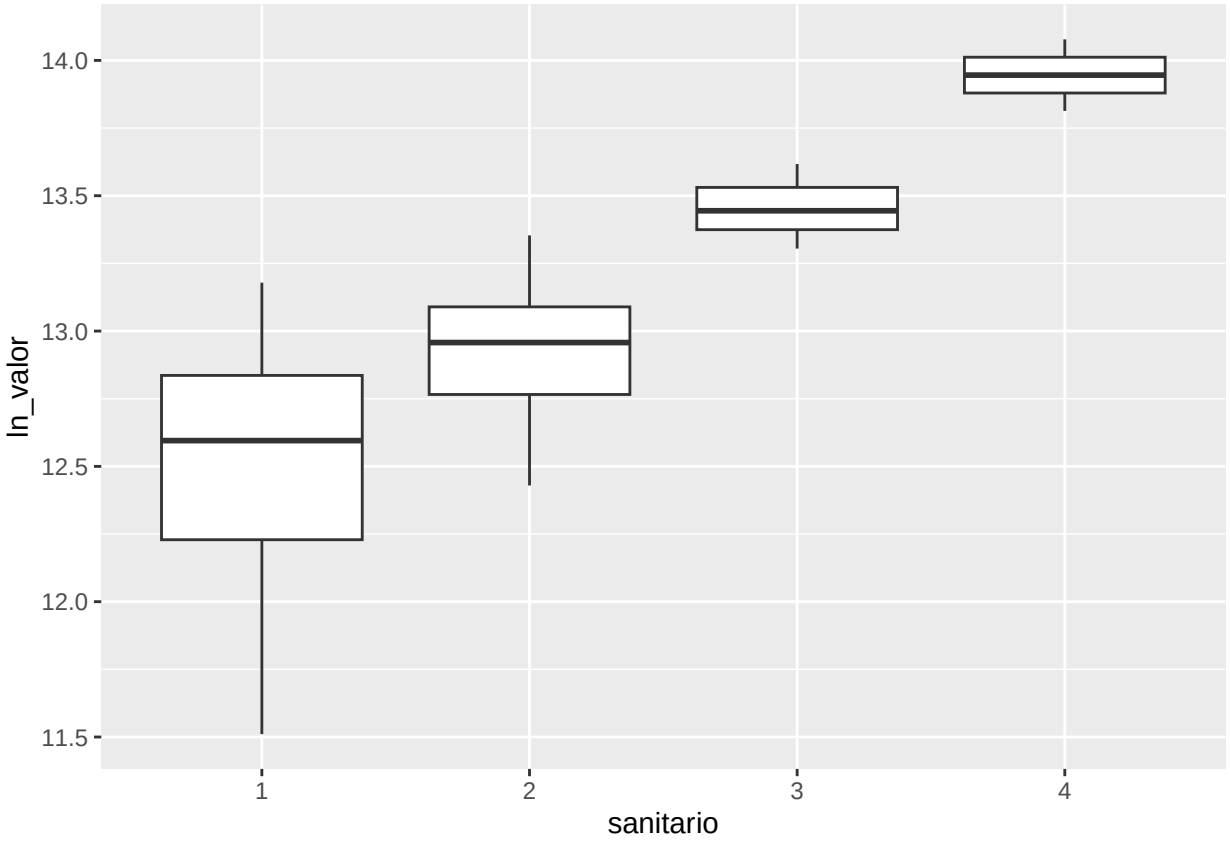
4. Interpretação econômica

- O número de quartos é um indicador importante para o valor do imóvel, mas ele sozinho não explica totalmente as variações. Outros fatores como área útil, localização, e padrão construtivo devem ser incluídos em uma análise mais detalhada.

sanitario x ln_valor

```
apartamento$sanitario <- as.factor(apartamento$sanitario)

ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = sanitario, y = ln_valor))
```



Com base no gráfico de boxplot que relaciona **sanitário** (número de banheiros) com **ln_valor** (logaritmo do valor de venda), podemos tirar as seguintes conclusões:

---

1. Relação positiva entre banheiros e valor

- Existe uma relação clara e positiva entre o número de banheiros e o valor do imóvel:
  - À medida que o número de banheiros aumenta, o logaritmo do valor de venda (**ln_valor**) também aumenta.
  - Isso reflete a maior atratividade e funcionalidade de imóveis com mais banheiros, geralmente associados a um maior padrão ou tamanho do imóvel.

---

2. Variação dentro dos grupos

- **1 banheiro**:
  - Apresenta a maior dispersão nos valores de **ln_valor**, o que sugere uma grande diversidade de características nos imóveis com apenas 1 banheiro.
- **2 banheiros**:
  - Mostra menor dispersão e uma mediana maior que os imóveis com 1 banheiro, indicando valores mais consistentes nesse grupo.

- **3 e 4 banheiros**:
  - Grupos mais homogêneos, com menor dispersão e valores de **ln_valor** consistentemente mais altos.
  - Imóveis com 4 banheiros têm os valores mais altos e uma dispersão muito baixa, indicando que esses imóveis são de padrão elevado.

---

3. Comparação entre grupos

- Há diferenças claras entre os valores medianos de **ln_valor** para cada grupo de banheiros, com aumento consistente conforme o número de banheiros cresce.
- Essa relação sugere que o número de banheiros é um indicador significativo para o valor do imóvel, especialmente para imóveis de maior padrão e tamanho.

---

4. Outliers

- O grupo de 1 banheiro apresenta um outlier (valor logarítmico baixo de **ln_valor**), que pode ser investigado como uma possível inconsistência ou uma característica muito diferente (exemplo: imóvel mal localizado ou em más condições).

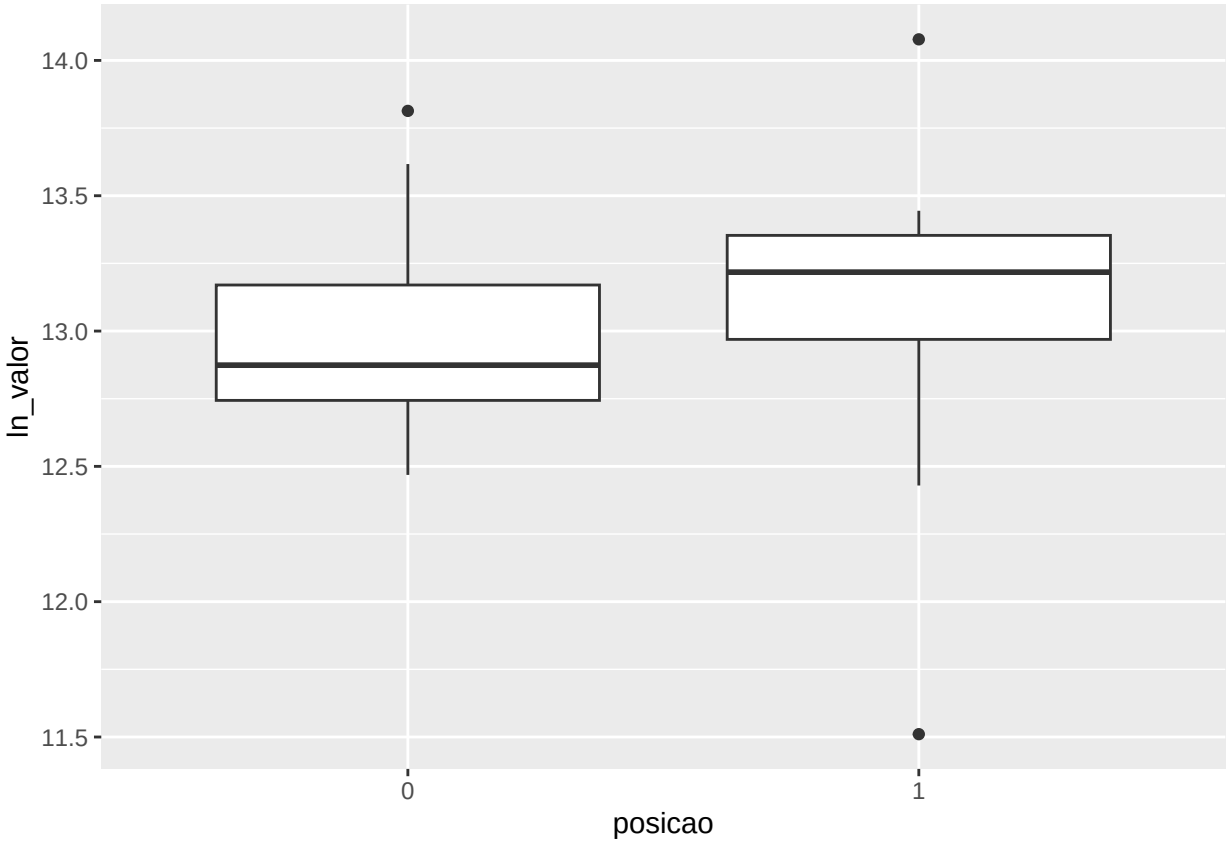
---

5. Interpretação econômica

- O número de banheiros é um fator importante que influencia positivamente o valor do imóvel, sendo também um indicativo de padrão e funcionalidade.
- Imóveis com mais banheiros tendem a ter maior valor agregado, possivelmente devido ao tamanho maior e à oferta de maior conforto para os moradores.

posicao x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = posicao, y = ln_valor))
```



Com base no boxplot que relaciona **posição** (posição do apartamento: `0` para não voltado para a rua e `1` para voltado para a rua) com **ln_valor** (logaritmo do valor de venda), as seguintes observações podem ser feitas:

---

1. Diferença entre os grupos

- Apartamentos voltados para a rua (**posição = 1**) apresentam uma mediana de **ln_valor** (valor logarítmico de venda) ligeiramente maior do que os apartamentos que não são voltados para a rua (**posição = 0**).
- Isso sugere que a posição voltada para a rua tem algum impacto no valor total do imóvel.

---

2. Dispersão dos valores

- **Posição = 0 (não voltado para a rua)**:
  - Apresenta maior dispersão indicando que há apartamentos não voltados para a rua que, ainda assim, têm valores de venda muito altos.
- **Posição = 1 (voltado para a rua)**:
  - Apresenta menor dispersão no geral, mas com um outlier abaixo de **ln_valor = 12**, sugerindo que alguns apartamentos voltados para a rua ainda têm valores relativamente baixos devido a outras características.

---

3. Comparação e sobreposição

- Apesar das medianas distintas, existe uma considerável sobreposição entre os valores de **ln_valor** para os dois grupos.
- Isso indica que a posição, por si só, não explica totalmente as diferenças no valor total do imóvel, mas pode ser um fator contribuidor.

---

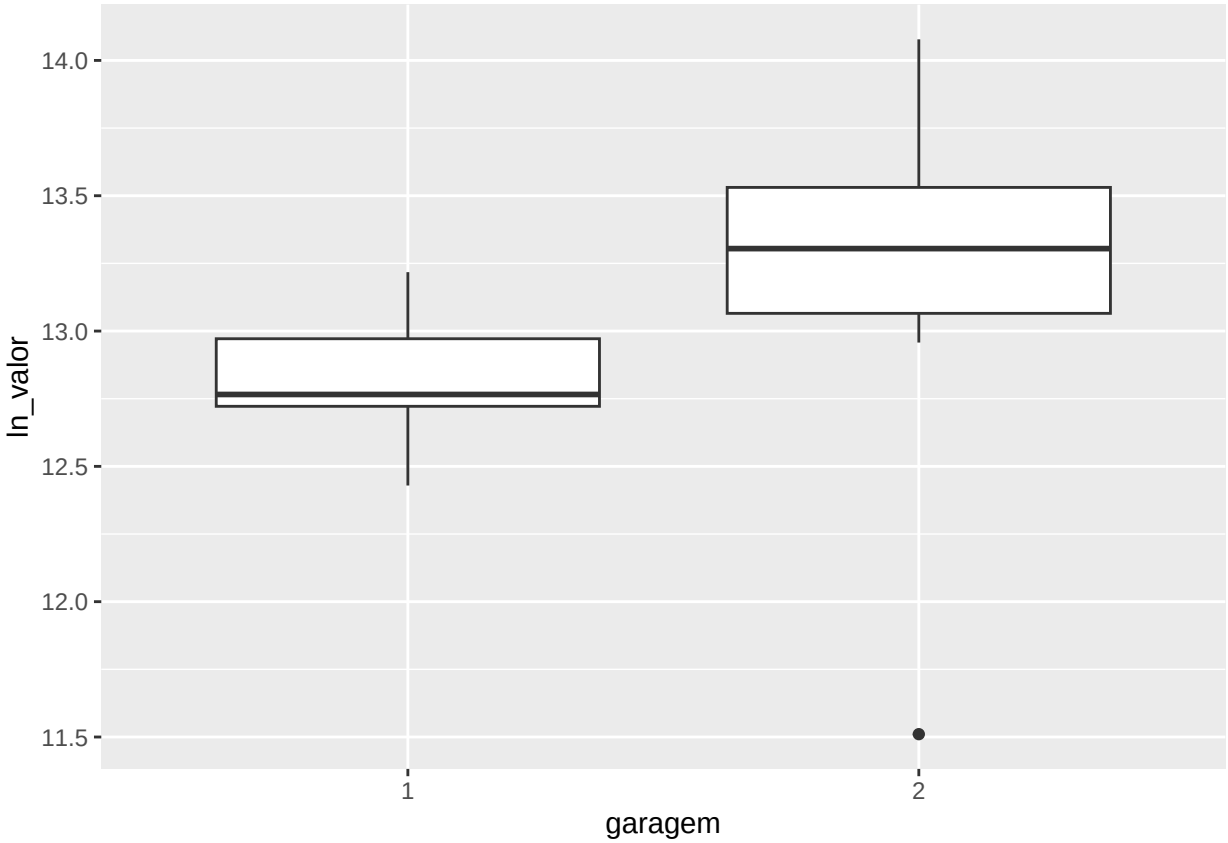
4. Interpretação econômica

- Apartamentos voltados para a rua tendem a ser percebidos como mais atrativos, podendo ter melhor iluminação, ventilação, ou uma vista mais agradável, o que impacta positivamente o valor total.
- No entanto, o impacto é moderado e outros fatores, como área, número de quartos, ou localização, podem ter maior relevância na determinação do preço.

garagem x ln_valor

```
apartamento$garagem <- as.factor(apartamento$garagem)

ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = garagem, y = ln_valor))
```



Com base no gráfico de boxplot que relaciona o número de vagas de garagem (**garagem**) com o logaritmo do valor de venda (**ln_valor**), podemos observar as seguintes conclusões:

---

1. Diferença entre os grupos

- **Imóveis com 2 vagas de garagem** possuem valores de **ln_valor** (log do valor de venda) mais altos em comparação com imóveis com 1 vaga de garagem.
- A mediana de **ln_valor** para imóveis com 2 vagas é visivelmente superior à dos imóveis com 1 vaga, indicando que mais vagas de garagem influenciam positivamente o valor de venda.

---

2. Dispersão e variabilidade

- **1 vaga de garagem**:
  - Apresenta menor dispersão nos valores de **ln_valor**, mas com valores concentrados em uma faixa mais baixa.
- **2 vagas de garagem**:
  - Maior dispersão nos valores de **ln_valor**, com um outlier em um valor muito baixo. Esse outlier pode representar um imóvel com características menos atrativas, apesar de possuir 2 vagas.

– A variabilidade maior pode estar relacionada a outras características dos imóveis com mais vagas, como maior área útil ou padrão construtivo elevado.

---

3. Impacto das vagas no valor de venda

- A presença de 2 vagas de garagem parece ser uma característica valorizada no mercado, contribuindo para preços mais altos.
- Isso pode ser explicado por fatores como maior comodidade para famílias maiores ou compradores que valorizam o espaço adicional.

---

4. Outliers

- O outlier no grupo de 2 vagas de garagem indica que nem todos os imóveis com mais vagas necessariamente têm valores elevados. Pode ser interessante investigar esse ponto para verificar inconsistências ou características adversas (ex.: localização ou padrão inferior).

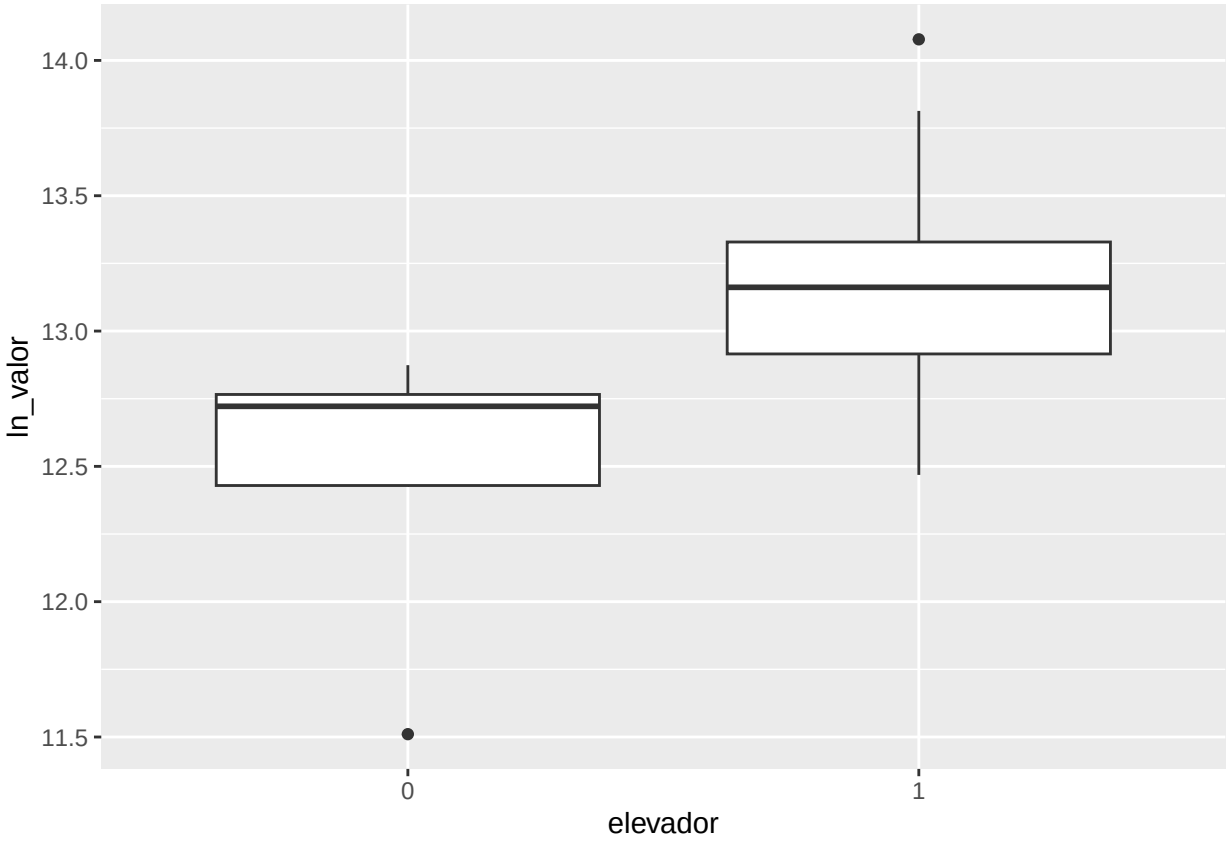
---

5. Comparação geral

- Há uma diferença clara entre os dois grupos, sugerindo que o número de vagas de garagem é um fator importante para determinar o valor do imóvel.
- Essa diferença pode ser explorada em modelos estatísticos para quantificar sua relevância em relação a outras variáveis, como área útil ou padrão construtivo.

elevador x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = elevador, y = ln_valor))
```



Com base no gráfico de boxplot que relaciona a presença de elevador (**elevador**, sendo `0` para ausência e `1` para presença) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Diferença entre os grupos

- Imóveis em prédios com **elevador** possuem valores medianos de **ln_valor** visivelmente superiores aos imóveis sem elevador.
- Essa diferença reflete a percepção de que a presença de elevador agrega valor aos imóveis, especialmente em prédios com múltiplos pavimentos, onde é um fator essencial de acessibilidade e conforto.

---

2. Dispersão dos valores

- **Prédios sem elevador (`elevador = 0`)**:
  - Apresentam menor dispersão nos valores de **ln_valor**, com a maioria dos valores concentrados em uma faixa inferior.
  - Há um outlier com valor muito baixo de **ln_valor**, indicando um imóvel de preço atípico ou com características adversas.
- **Prédios com elevador (`elevador = 1`)**:
  - Maior dispersão, incluindo um outlier com valor de **ln_valor** muito alto, possivelmente um imóvel de alto padrão ou características únicas.

– A faixa de valores superiores reforça que o elevador está frequentemente associado a imóveis de padrão mais elevado.

---

3. Impacto da presença de elevador

- A diferença entre as medianas mostra que a presença de elevador é um fator valorizado no mercado imobiliário, principalmente em prédios mais modernos ou com maior número de pavimentos.
- A ausência de elevador está associada a imóveis de menor valor, provavelmente limitados a edifícios de poucos andares ou construções mais antigas.

---

4. Outliers

- O outlier no grupo sem elevador pode representar um imóvel com características únicas além da falta de elevador (exemplo: localização ou pequena área).
- O outlier no grupo com elevador pode ser um imóvel de altíssimo padrão ou localização excepcional.

---

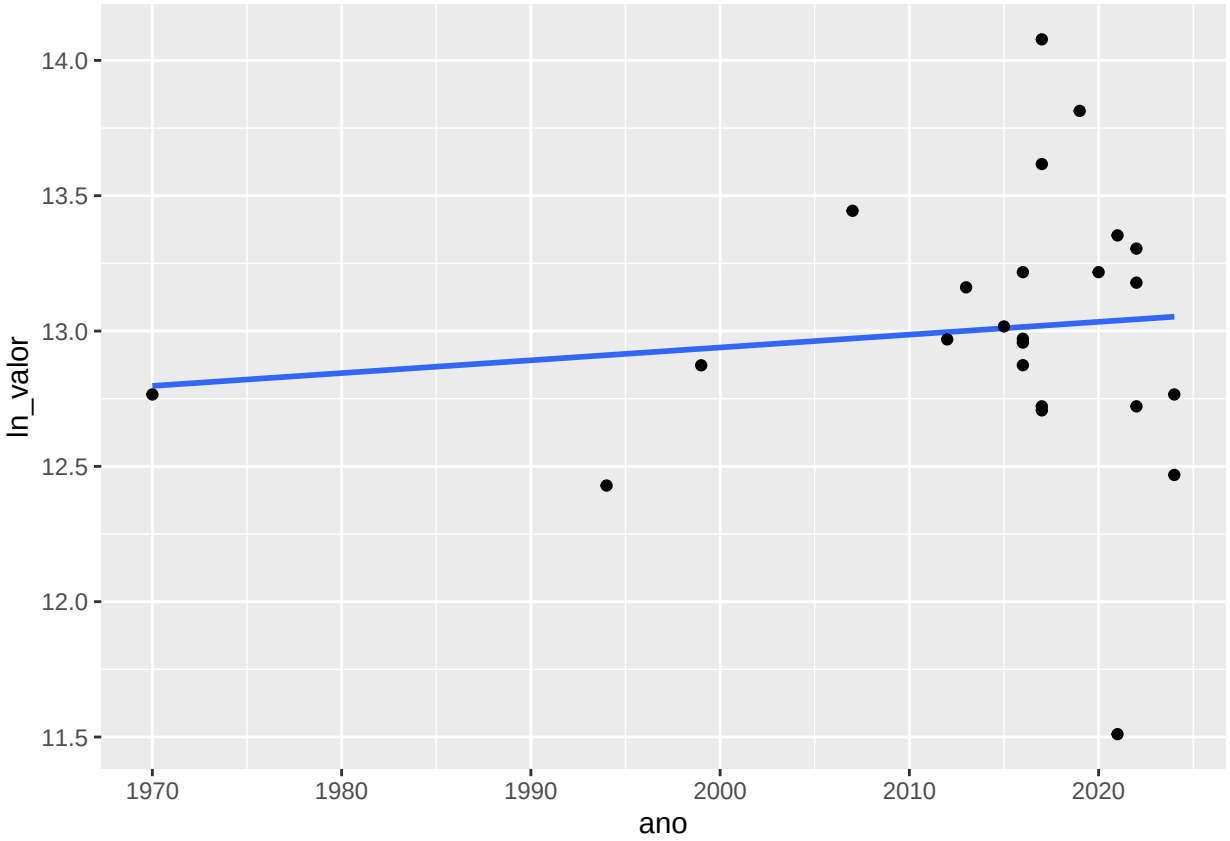
5. Interpretação econômica

- A presença de elevador agrega valor significativo aos imóveis, especialmente em edifícios que atendem famílias ou pessoas que necessitam de acessibilidade.
- No entanto, outros fatores como **área útil**, **padrão construtivo** e **localização** provavelmente interagem com a presença de elevador para determinar o valor final do imóvel.

ano x ln_valor

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = ano, y = ln_valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = ano, y = ln_valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico de dispersão com uma linha de tendência linear ajustada, que relaciona o ano de construção (**ano**) no eixo X e o logaritmo do valor de venda (**ln_valor**) no eixo Y, podemos tirar as seguintes conclusões:

---

1. Relação entre ano e ln_valor

- Há uma **tendência positiva fraca** entre o ano de construção e o logaritmo do valor de venda:
  - Imóveis mais novos (construídos em anos mais recentes) tendem a ter valores de venda levemente mais altos em logaritmo.
  - Isso pode refletir o fato de que imóveis mais novos geralmente apresentam padrões construtivos modernos e menores necessidades de manutenção, tornando-os mais atraentes.

---

2. Dispersão nos valores

- **Imóveis construídos antes de 2000**:
  - Mostram maior variabilidade nos valores de **ln_valor**, incluindo outliers com valores muito baixos. Isso pode refletir diferenças significativas de qualidade ou localização desses imóveis.
- **Imóveis construídos após 2000**:

- Há maior concentração de valores em uma faixa mais alta de **ln_valor**, indicando que os imóveis modernos tendem a ter valores mais consistentes.
- Mesmo assim, alguns imóveis construídos recentemente apresentam valores baixos, possivelmente devido a localização desfavorável ou características específicas.

---

3. Linha de tendência linear

- A inclinação positiva da linha de tendência é leve, indicando que o ano de construção sozinho não explica grandes variações no valor do imóvel.
- Isso sugere que outras variáveis, como localização, área útil, ou padrão construtivo, têm papel mais significativo na determinação do preço.

---

4. Outliers

- Há outliers nos dois extremos:
  - **Ano mais antigo com alto ln_valor**: Pode indicar um imóvel antigo em localização privilegiada ou com reformas significativas.
  - **Ano recente com baixo ln_valor**: Pode representar um imóvel de baixo padrão ou localizado em uma área menos valorizada.

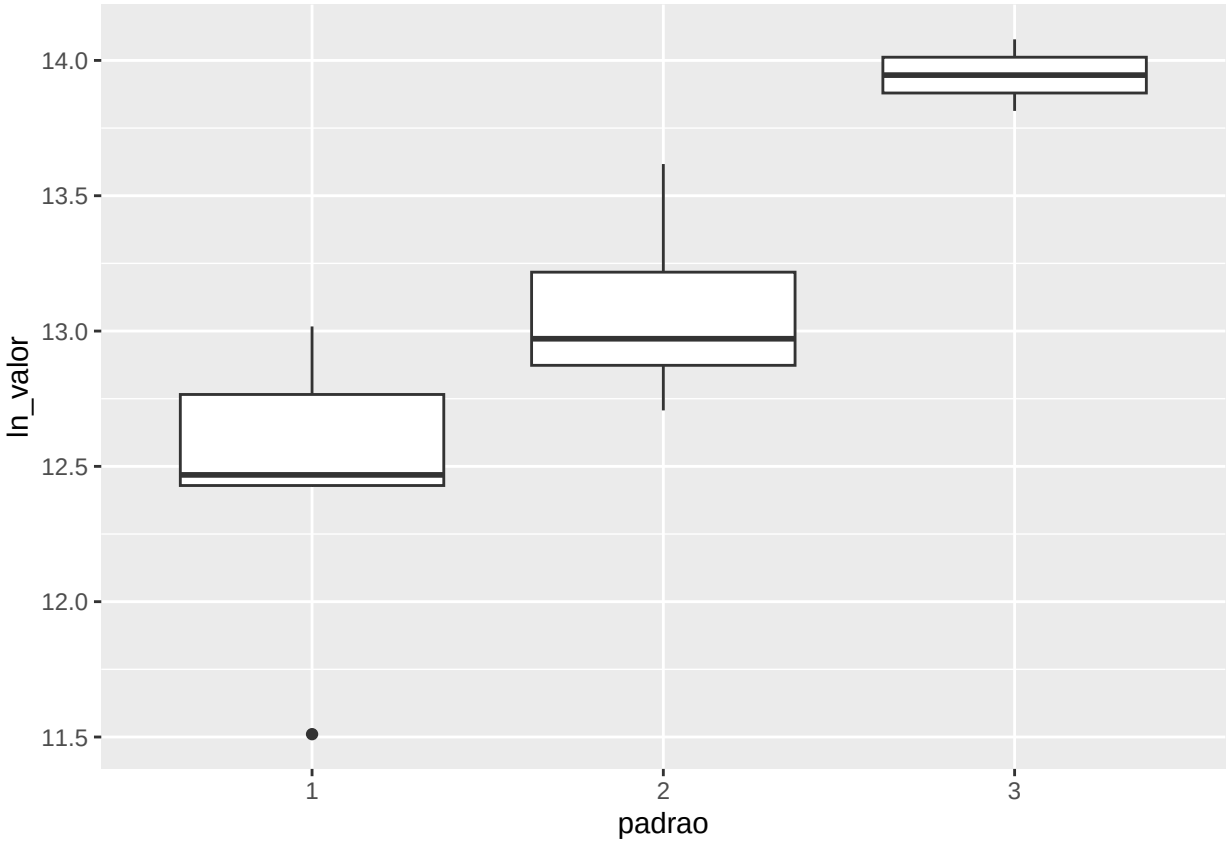
---

5. Interpretação econômica

- Imóveis mais recentes geralmente possuem características que aumentam seu valor, mas o ano de construção sozinho não é um preditor suficientemente forte.
- A valorização de imóveis também depende de fatores como manutenção, reformas, e localização, o que explica a dispersão observada.

padrao x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = padrao, y = ln_valor))
```



Com base no boxplot que relaciona o padrão construtivo (**padrao**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Diferença entre os grupos

- O valor de **ln_valor** aumenta consistentemente à medida que o padrão construtivo do imóvel sobe:
  - **Padrão 1 (baixo)**: Os imóveis têm os menores valores medianos de **ln_valor**.
  - **Padrão 2 (normal)**: Apresentam valores medianos de **ln_valor** superiores ao padrão 1, mas com maior dispersão.
  - **Padrão 3 (alto)**: Possuem os maiores valores medianos de **ln_valor**, com pouca dispersão.

---

2. Dispersão dos valores

- **Padrão 1 (baixo)**:
  - Apresenta menor dispersão, mas contém um outlier com valor muito baixo de **ln_valor**, possivelmente representando um imóvel com características adversas.
- **Padrão 2 (normal)**:
  - Maior dispersão nos valores de **ln_valor**, refletindo a diversidade de imóveis nesse grupo (possivelmente variando em localização, área útil, ou outras características).

- **Padrão 3 (alto)**:
  - Valores concentrados em uma faixa superior, com pouca variabilidade, indicando que esses imóveis têm características mais uniformes e de alta qualidade.

---

3. Impacto do padrão no valor

- O padrão construtivo é um fator fortemente associado ao valor do imóvel, com aumento claro de **ln_valor** conforme o padrão sobe.
- Isso reflete a percepção de qualidade e conforto dos imóveis: padrões mais altos tendem a incluir acabamentos de luxo, materiais de melhor qualidade e maior valorização no mercado.

---

4. Outlier

- No grupo **Padrão 1 (baixo)**, há um outlier com valor de **ln_valor** muito baixo. Esse imóvel pode ser investigado para verificar se há inconsistências ou características específicas que o tornam uma exceção.

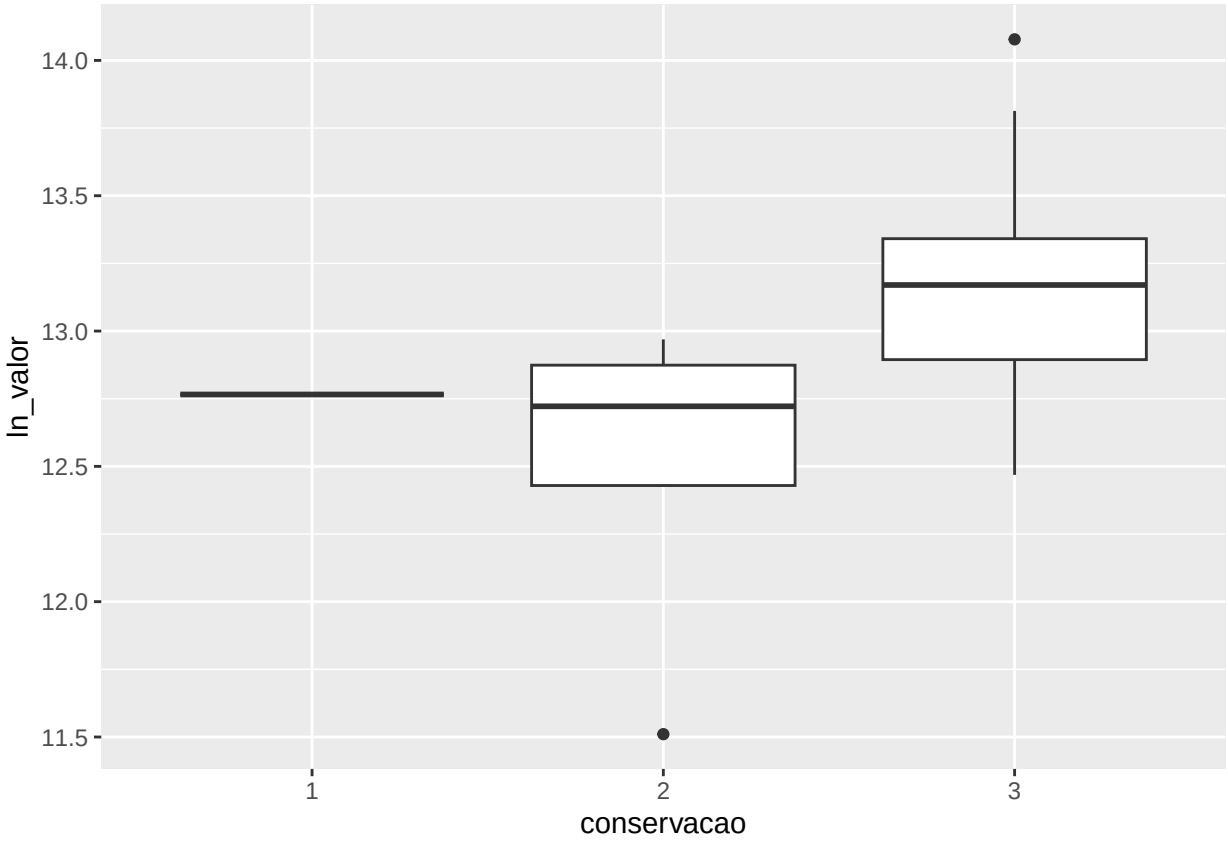
---

5. Comparação geral

- A separação entre os grupos é bem marcada, com pouca sobreposição entre os padrões 1 e 3.
- Isso indica que o padrão construtivo é uma variável importante para diferenciar os valores de venda dos imóveis.

conservacao x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = conservacao, y = ln_valor))
```



Com base no boxplot que relaciona o estado de conservação (**conservacao**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Diferença entre os grupos

- **Conservação 1 (ruim)**:
  - Poucos imóveis estão nesse grupo, com valores concentrados em uma faixa baixa de **ln_valor**.
- **Conservação 2 (bom)**:
  - Valores de **ln_valor** são mais altos do que no grupo de conservação ruim, mas com maior dispersão, incluindo um outlier com valor muito baixo.
- **Conservação 3 (excelente)**:
  - Possui os maiores valores medianos de **ln_valor** e uma dispersão moderada, com um outlier para valores muito altos.
  - Imóveis nesse grupo têm claramente os preços mais altos, reforçando que o estado de conservação influencia significativamente o valor do imóvel.

---

2. Impacto da conservação no valor

- A tendência é clara: imóveis em melhor estado de conservação (valores mais altos de **conservacao**) possuem valores de venda mais altos.
- Isso reflete a atratividade e a percepção de qualidade do imóvel no mercado: imóveis bem conservados requerem menos reparos e são mais valorizados.

---

3. Dispersão e outliers

- **Conservação 2 (bom)**:
  - Inclui um outlier com **ln_valor** muito baixo. Esse imóvel pode ter características adversas (ex.: localização ruim ou padrão baixo) que compensam o bom estado de conservação.
- **Conservação 3 (excelente)**:
  - Possui um outlier com **ln_valor** muito alto, indicando um imóvel de padrão elevado ou localização excepcional.

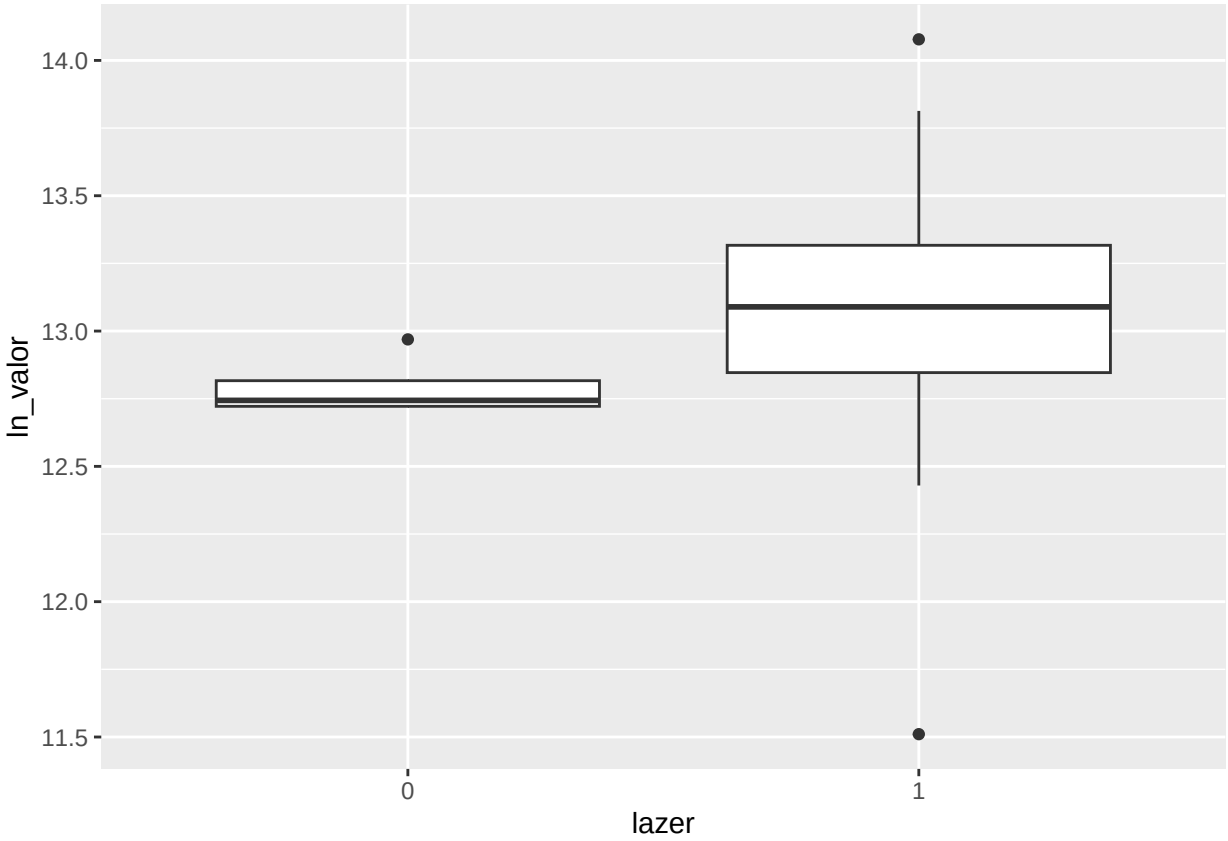
---

4. Comparação entre os grupos

- Há uma diferença clara e consistente entre os grupos de conservação, com uma hierarquia bem definida (conservação ruim < bom < excelente).
- Essa relação é menos sobreposta em comparação a outras variáveis analisadas, sugerindo que o estado de conservação é um indicador importante e bem diferenciado para o valor dos imóveis.

lazer x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = lazer, y = ln_valor))
```



Com base no boxplot que relaciona a presença de área de lazer no prédio (**lazer**, sendo `0` para ausência e `1` para presença) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Diferença entre os grupos

- **Prédios com área de lazer (`lazer = 1`)**:
  - Apresentam valores medianos de **ln_valor** significativamente maiores do que prédios sem área de lazer.
  - Isso reflete que a presença de área de lazer é uma característica valorizada no mercado imobiliário, agregando valor ao imóvel.
- **Prédios sem área de lazer (`lazer = 0`)**:
  - Apresentam valores mais baixos e menor dispersão em **ln_valor**, com valores concentrados em uma faixa inferior.

---

2. Dispersão dos valores

- **Prédios com área de lazer (`lazer = 1`)**:

  - Apresentam maior dispersão, com valores que variam de **ln_valor** moderado a muito alto. Isso pode indicar diversidade nas características dos prédios com lazer (por exemplo, padrão, localização ou qualidade das áreas de lazer).
  - Há um outlier com valor de **ln_valor** muito baixo, possivelmente representando um imóvel com área de lazer, mas de padrão inferior ou em uma localização menos valorizada.

- **Prédios sem área de lazer (`lazer = 0`)**:
  - Apresentam menor dispersão e um valor outlier mais alto, que pode ser um imóvel sem área de lazer, mas com outras características de destaque (como localização privilegiada).

---

3. Impacto da área de lazer

- A presença de área de lazer é um fator importante para a valorização do imóvel, contribuindo para preços mais altos.
- Imóveis com área de lazer geralmente estão associados a padrões construtivos mais elevados e maior conforto, justificando a diferença nos valores de **ln_valor**.

---

4. Outliers

- **No grupo `lazer = 0`**, o outlier indica um imóvel que, apesar de não ter área de lazer, possui características que aumentam seu valor (ex.: localização, tamanho, ou padrão construtivo).
- **No grupo `lazer = 1`**, o outlier inferior indica que a área de lazer sozinha não garante um valor alto, especialmente se o imóvel tiver outras características desfavoráveis.

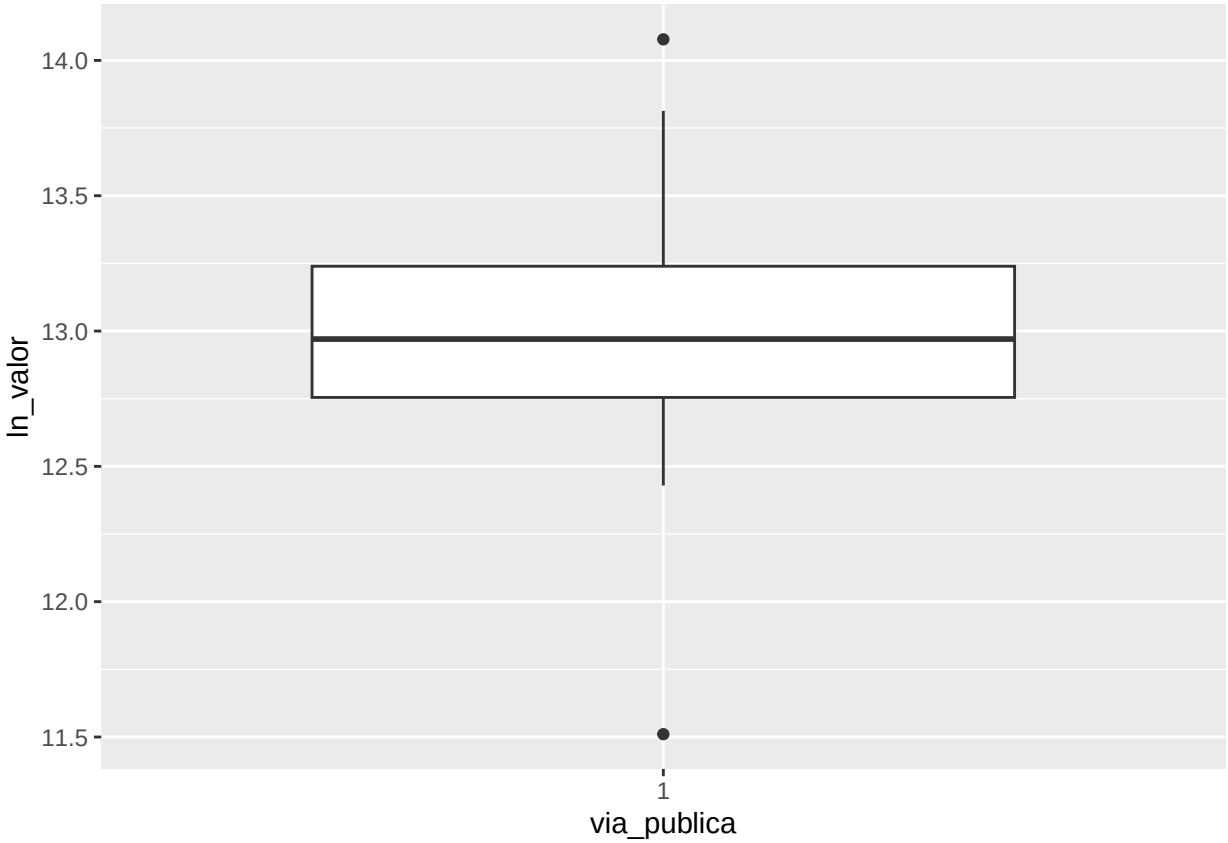
---

5. Comparação geral

- A diferença clara entre os dois grupos sugere que a presença de área de lazer é uma característica valorizada no mercado, mas sua influência também pode depender de outros fatores.

via_publica x ln_valor

```
ggplot(data = apartamento) +
  geom_boxplot(mapping = aes(x = via_publica, y = ln_valor))
```



Com base no boxplot que relaciona a pavimentação da via pública (**via_publica**, sendo `1` para via pavimentada) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Dados homogêneos

- Como todos os imóveis no conjunto de dados têm **via_publica = 1** (via pavimentada), não há distinção entre grupos para análise comparativa.
- Este gráfico reflete a distribuição dos valores de **ln_valor** apenas para imóveis localizados em vias pavimentadas.

---

2. Dispersão dos valores

- A dispersão dos valores de **ln_valor** é moderada, com a maioria dos imóveis concentrados ao redor da mediana.
- Existem dois outliers:
  - **Outlier superior**: Representa um imóvel com **ln_valor** muito alto, provavelmente devido a características diferenciadas como localização privilegiada, padrão construtivo elevado ou amenidades extras.
  - **Outlier inferior**: Representa um imóvel com **ln_valor** muito baixo, possivelmente devido a características adversas como padrão baixo, pequena área útil ou localização desfavorável.

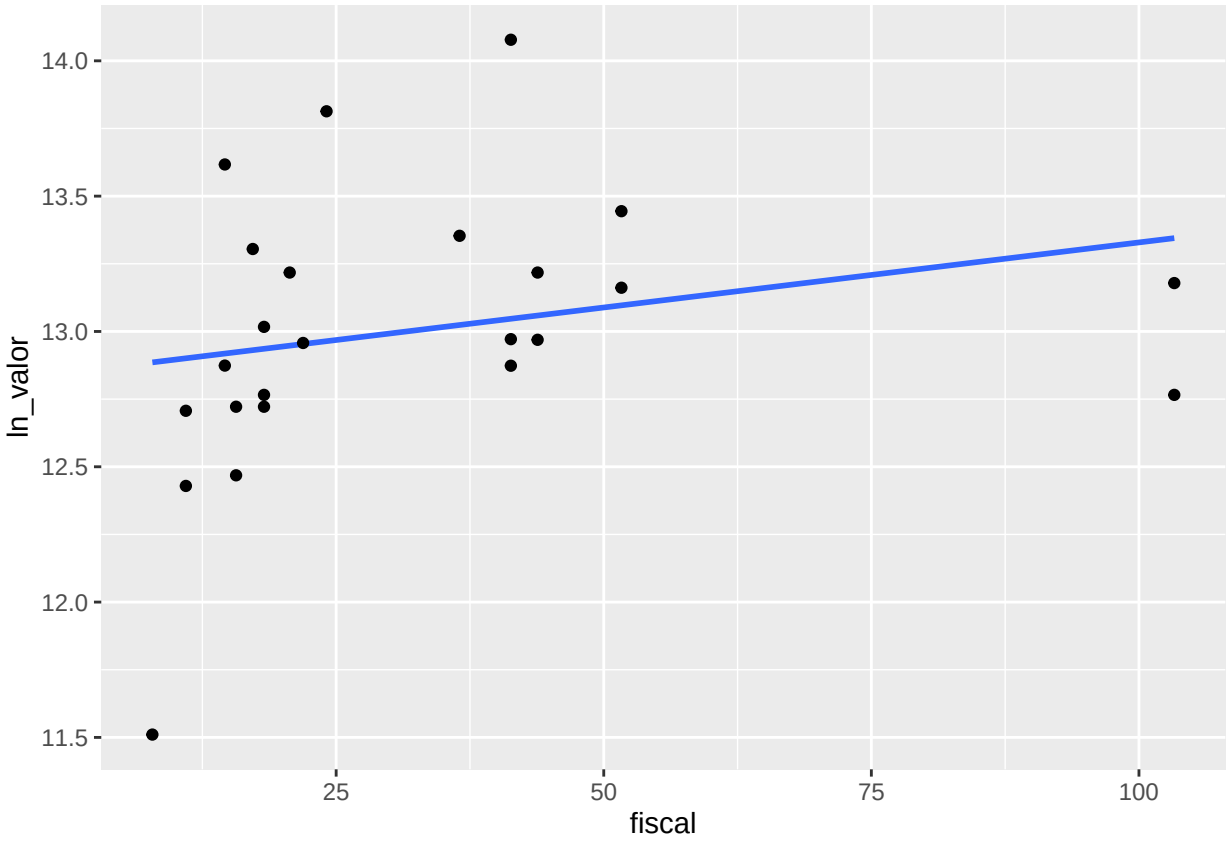
3. Impacto da pavimentação

- Como não há imóveis em vias não pavimentadas no conjunto de dados, não é possível avaliar diretamente o impacto da pavimentação da via pública sobre o valor do imóvel.
- No entanto, a pavimentação é geralmente um fator valorizado no mercado imobiliário, podendo influenciar positivamente o preço dos imóveis.

fiscal x ln_valor

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = fiscal, y = ln_valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = fiscal, y = ln_valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico de dispersão com a linha de tendência linear ajustada, que relaciona o valor fiscal da área (**fiscal**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

1. **Relação positiva moderada**

- Existe uma **tendência positiva** entre o valor fiscal e o logaritmo do valor de venda:
  - Imóveis localizados em áreas com maior valor fiscal tendem a apresentar valores mais altos de **ln_valor**.
  - Isso reflete a valorização das regiões onde o valor fiscal é mais alto, geralmente associado a melhores infraestruturas, localização privilegiada ou alta demanda.

---

2. **Dispersão dos valores**

- Os valores de **ln_valor** mostram dispersão em quase toda a faixa de valores fiscais:
  - Para valores fiscais baixos (< 25), há maior dispersão, incluindo outliers com valores de **ln_valor** muito baixos e muito altos.
  - Para valores fiscais altos (> 75), a dispersão diminui, com valores de **ln_valor** mais concentrados na faixa superior.

---

3. **Impacto do valor fiscal**

- O valor fiscal é um indicador importante para a valorização do imóvel, mas a dispersão nos dados sugere que ele não é o único fator determinante.
- Outras variáveis, como área útil, padrão construtivo, e presença de amenidades, também desempenham papéis significativos.

---

4. **Outliers**

- Há outliers em ambas as extremidades:
  - **Outlier com valor fiscal baixo e ln_valor baixo**: Pode indicar um imóvel em localização menos favorecida ou de padrão inferior.
  - **Outlier com valor fiscal alto e ln_valor baixo**: Representa uma discrepância que pode ser analisada para verificar possíveis inconsistências ou características específicas.

---

5. **Interpretação econômica**

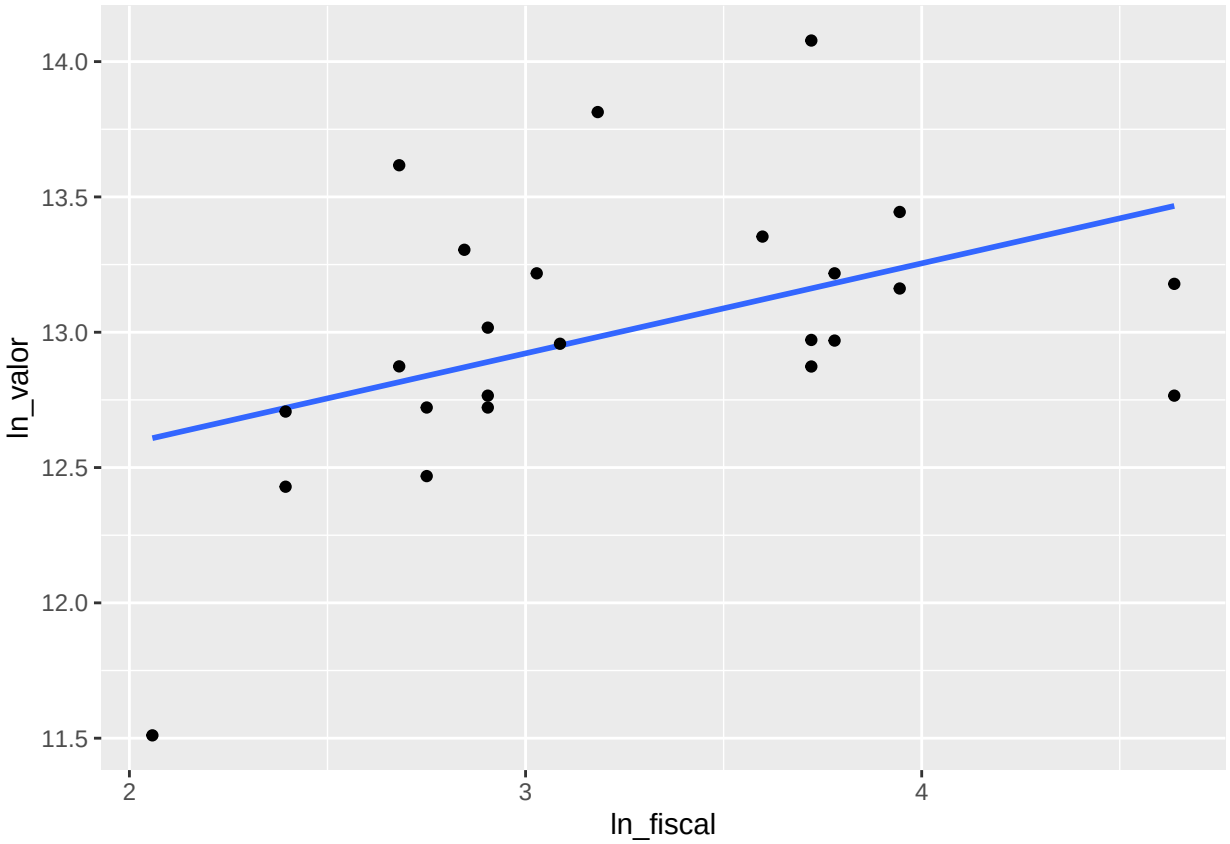
- Áreas com maior valor fiscal geralmente correspondem a regiões mais valorizadas, o que se reflete nos preços de venda mais altos.
- No entanto, a variabilidade indica que o impacto do valor fiscal é moderado e precisa ser analisado em conjunto com outros atributos.

transformando fiscal = ln(fiscal)

```r
apartamento$ln_fiscal <- log(apartamento$fiscal)

ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = ln_fiscal, y = ln_valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = ln_fiscal, y = ln_valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico de dispersão com linha de tendência linear ajustada, que relaciona o logaritmo do valor fiscal (**ln_fiscal**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Relação positiva clara

- Há uma **relação linear positiva moderada** entre **ln_fiscal** e **ln_valor**:
  - Conforme o logaritmo do valor fiscal aumenta, o logaritmo do valor de venda também tende a aumentar.
  - Isso reforça que imóveis localizados em áreas com maior valor fiscal são, geralmente, mais valorizados.

---

2. Dispersão dos valores

- A dispersão dos pontos ao longo da linha de tendência sugere que **ln_fiscal** influencia **ln_valor**, mas não é o único determinante:
  - Há variabilidade significativa nos valores de **ln_valor** para os mesmos níveis de **ln_fiscal**, indicando que outros fatores (como área útil, padrão, ou conservação) também desempenham papéis importantes.

---

3. Outliers

- **Outlier inferior**: Existe um imóvel com **ln_valor** muito baixo em relação ao **ln_fiscal**, indicando características adversas que afetam o preço (ex.: localização específica, padrão baixo ou tamanho pequeno).
- **Outlier superior**: Um imóvel apresenta **ln_valor** muito alto em relação ao **ln_fiscal**, possivelmente devido a características diferenciadas como alta qualidade construtiva ou localização privilegiada.

---

4. Impacto de ln_fiscal no ln_valor

- A relação positiva entre **ln_fiscal** e **ln_valor** é consistente e sugere que o valor fiscal é um indicador de valorização do imóvel.
- A transformação logarítmica ajuda a estabilizar a variabilidade, permitindo uma análise mais clara da relação linear entre as variáveis.

---

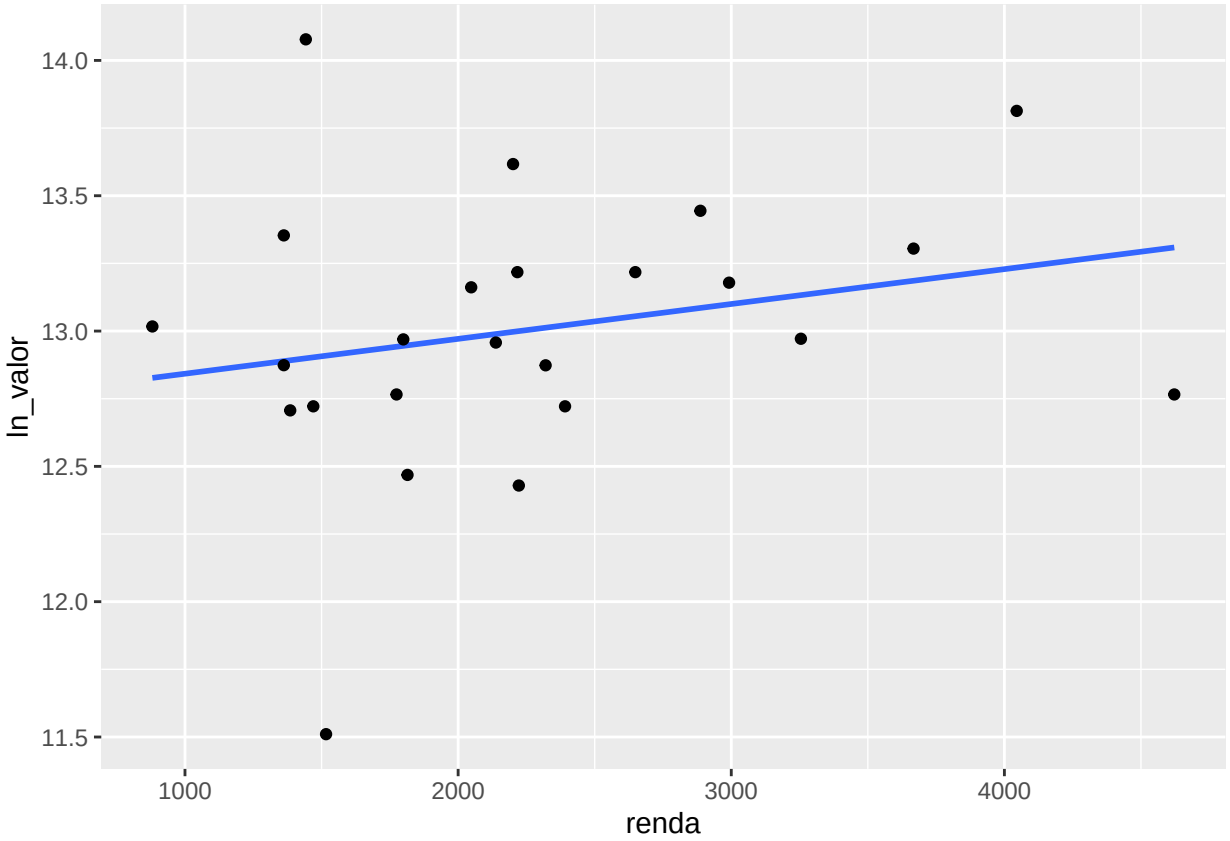
5. Comparação com análise anterior (fiscal x ln_valor)

- A relação logarítmica suaviza a dispersão observada no gráfico anterior, evidenciando uma relação linear mais consistente.
- Isso indica que o uso do logaritmo é apropriado para modelar a relação entre o valor fiscal e o valor de venda.

renda x ln_valor

```
ggplot(data = apartamento) +
  geom_smooth(mapping = aes(x = renda, y = ln_valor), method = "lm", se = FALSE) +
  geom_point(mapping = aes(x = renda, y = ln_valor))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Com base no gráfico de dispersão com a linha de tendência linear ajustada, que relaciona a renda média do bairro (**renda**) com o logaritmo do valor de venda (**ln_valor**), podemos tirar as seguintes conclusões:

---

1. Relação positiva moderada

- Existe uma **relação positiva fraca a moderada** entre a renda média do bairro e o logaritmo do valor de venda:
  - Imóveis em bairros com maior renda média tendem a ter valores de venda mais altos (em logaritmo).
  - Isso reflete a valorização de bairros com melhores condições socioeconômicas, o que frequentemente se traduz em melhor infraestrutura, localização privilegiada ou maior demanda.

---

2. Dispersão dos valores

- Há variabilidade significativa em **ln_valor** dentro de cada faixa de renda:
  - Mesmo em bairros de renda média alta, existem imóveis com valores relativamente baixos de **ln_valor**, indicando que outros fatores (como padrão construtivo, área útil ou conservação) também influenciam os preços.
  - Nos bairros de renda média mais baixa, a dispersão de valores é maior, sugerindo uma diversidade maior de tipos de imóveis.

---

3. Impacto da renda média

- Embora a renda média tenha uma relação positiva com o valor do imóvel, a dispersão indica que ela sozinha não é suficiente para prever o preço com alta precisão.
- A linha de tendência mostra um crescimento suave, sugerindo que o impacto da renda é moderado e deve ser combinado com outras variáveis para uma explicação mais robusta.

---

4. Outliers

- Existem outliers visíveis:
  - **Outlier com baixa renda e baixo ln_valor**: Pode representar um imóvel em uma região desfavorecida, com características adversas como padrão baixo ou pequena área útil.
  - **Outlier com alta renda e alto ln_valor**: Representa um imóvel em uma região valorizada e com características que justificam o preço elevado.

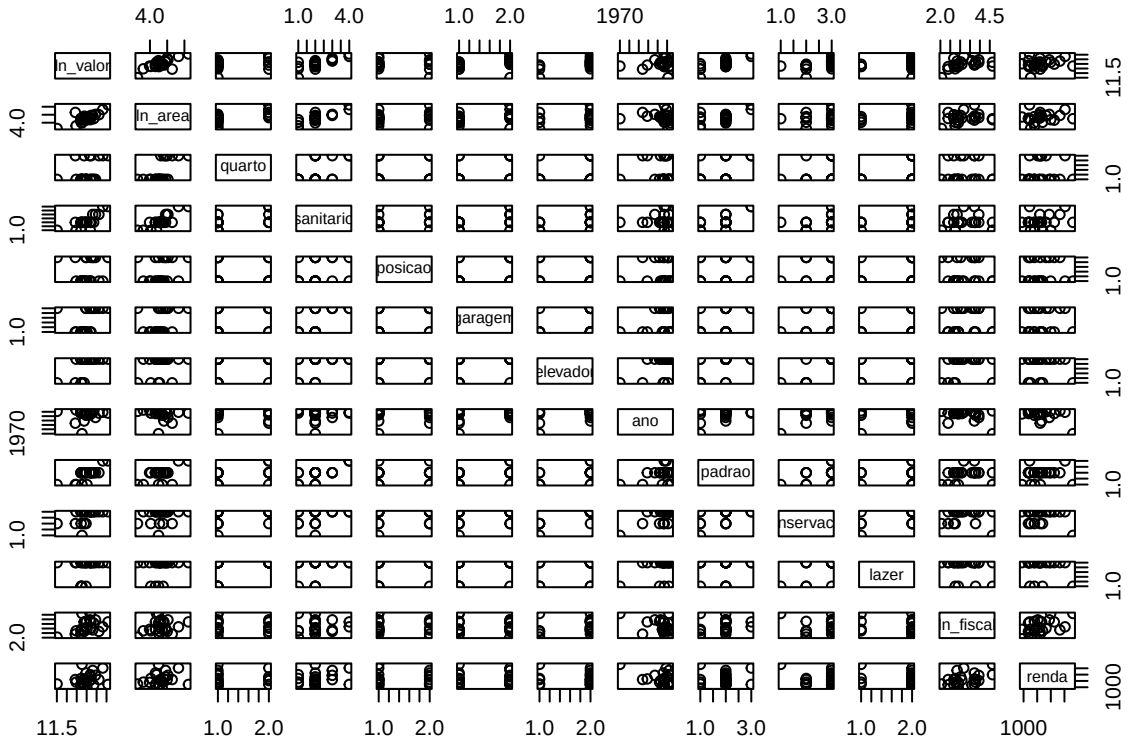
---

5. Interpretação econômica

- A renda média do bairro influencia positivamente o valor do imóvel, mas outros fatores, como características específicas do imóvel e sua localização dentro do bairro, também desempenham papéis importantes.

pares e correlacoes

```
pairs(~ ln_valor + ln_area + quarto + sanitario + posicao + garagem + elevador + ano + padrao + conserva
```




```
cor_matrix <- cor(apartamento[, c("ln_valor", "ln_area", "ano", "ln_fiscal", "renda")])

cor_matrix
```

```
##            ln_valor    ln_area        ano  ln_fiscal      renda
## ln_valor  1.0000000  0.7931215  0.1098123  0.4467824  0.2335210
## ln_area   0.7931215  1.0000000 -0.1977835  0.2403424  0.1922952
## ano       0.1098123 -0.1977835  1.0000000 -0.3900204 -0.4468904
## ln_fiscal 0.4467824  0.2403424 -0.3900204  1.0000000  0.4556536
## renda     0.2335210  0.1922952 -0.4468904  0.4556536  1.0000000
```

Com base na análise visual da matriz de pares e na matriz de correlação calculada, podemos tirar as seguintes conclusões:

---

1. Relação entre `ln_valor` e outras variáveis

- **`ln_area` ($r = 0.793$)**:
  - Existe uma forte correlação positiva entre o logaritmo da área e o logaritmo do valor de venda. Isso indica que, conforme a área aumenta, o valor de venda tende a aumentar significativamente.
- **`ln_fiscal` ($r = 0.447$)**:

  - Correlação positiva moderada. Áreas com maior valor fiscal tendem a ter imóveis mais valorizados, mas a relação não é tão forte quanto a de `ln_area`.

- `renda` **(r = 0.234)**:
  - Correlação fraca, mas positiva. Bairros com maior renda média têm imóveis ligeiramente mais valorizados.

- `ano` **(r = 0.110)**:
  - Correlação muito fraca. O ano de construção não está fortemente associado ao valor de venda.

---

2. Relação entre `ln_area` e outras variáveis

- `ln_area` **e** `ln_valor` **(r = 0.793)**:
  - A relação entre área e valor de venda é a mais forte na matriz, mostrando a importância da área útil na precificação do imóvel.

- `ln_area` **e** `ano` **(r = -0.198)**:
  - Correlação fraca e negativa, indicando que imóveis mais novos tendem a ser ligeiramente menores em área.

- `ln_area` **e** `ln_fiscal` **(r = 0.240)**:
  - Correlação fraca e positiva. Imóveis em áreas com maior valor fiscal tendem a ter áreas maiores.

---

3. Relação entre `ln_fiscal` e outras variáveis

- `ln_fiscal` **e** `ln_valor` **(r = 0.447)**:
  - Correlação positiva moderada. Áreas com maior valor fiscal estão associadas a valores de venda mais altos.

- `ln_fiscal` **e** `renda` **(r = 0.456)**:
  - Correlação moderada positiva. Bairros com maior valor fiscal geralmente têm maior renda média, sugerindo um padrão socioeconômico mais elevado.

- `ln_fiscal` **e** `ano` **(r = -0.390)**:
  - Correlação negativa moderada. Imóveis em áreas de maior valor fiscal tendem a ser mais antigos.

---

4. Relação entre `renda` e outras variáveis

- `renda` **e** `ln_valor` **(r = 0.234)**:
  - Correlação positiva fraca. Bairros com maior renda média têm valores de venda um pouco mais altos.

- `renda` **e** `ln_fiscal` **(r = 0.456)**:
  - Correlação moderada positiva, sugerindo que bairros com maior renda têm maior valor fiscal.

- **`renda` e `ano` (r = -0.447)**:
  - Correlação negativa moderada. Bairros com maior renda média tendem a ter imóveis mais antigos, possivelmente devido à maior consolidação desses bairros.

---

5. Relação entre `ano` e outras variáveis

- **`ano` e `ln_valor` (r = 0.110)**:
  - Correlação muito fraca e positiva. O ano de construção não impacta diretamente no valor de venda.
- **`ano` e `ln_fiscal` (r = -0.390)**:
  - Correlação negativa moderada. Áreas de maior valor fiscal tendem a ter imóveis mais antigos.
- **`ano` e `renda` (r = -0.447)**:
  - Correlação negativa moderada. Bairros com maior renda média geralmente têm imóveis mais antigos.

---

6. Considerações gerais

- **Variável mais influente para `ln_valor`:**
  - **`ln_area`** é a variável mais fortemente correlacionada com o valor de venda.
- **Correlação entre variáveis independentes:**
  - A correlação moderada entre **`ln_fiscal`** e **`renda`** (r = 0.456) indica que há alguma relação entre esses atributos, mas eles ainda fornecem informações diferentes.
- **Fracas relações com `ano`:**
  - O ano de construção tem pouca influência direta no valor de venda, mas está relacionado a outras variáveis, como o valor fiscal e a renda média.

1ª regressao linear

```
modelo_01 <- lm(ln_valor ~ ln_area + quarto + sanitario + posicao + garagem + elevador + ano + padrao +

modelo_01
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     garagem + elevador + ano + padrao + conservacao + lazer +
##     ln_fiscal + renda, data = apartamento)
##
## Coefficients:
##  (Intercept)       ln_area       quarto3    sanitario2    sanitario3
```

```
##   -4.810e+01     1.277e+00     4.611e-02     9.345e-02     3.898e-01
##   sanitario4      posicao1      garagem2     elevador1           ano
##    1.368e-01    -1.321e-01    -6.439e-03    -1.018e-01     2.746e-02
##      padrao2       padrao3  conservacao2  conservacao3        lazer1
##    8.430e-02            NA    -7.627e-01    -6.641e-01    -3.843e-03
##    ln_fiscal         renda
##    2.867e-01    -1.579e-05
```

```
summary(modelo_01)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     garagem + elevador + ano + padrao + conservacao + lazer +
##     ln_fiscal + renda, data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22839 -0.06003  0.00000  0.06410  0.25813
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.810e+01  2.399e+01  -2.005   0.0799 .
## ln_area       1.277e+00  4.101e-01   3.113   0.0144 *
## quarto3       4.611e-02  1.604e-01   0.287   0.7810
## sanitario2    9.345e-02  2.406e-01   0.388   0.7079
## sanitario3    3.898e-01  3.619e-01   1.077   0.3128
## sanitario4    1.368e-01  4.604e-01   0.297   0.7739
## posicao1     -1.321e-01  1.203e-01  -1.098   0.3041
## garagem2     -6.439e-03  1.436e-01  -0.045   0.9653
## elevador1    -1.018e-01  3.495e-01  -0.291   0.7781
## ano           2.746e-02  1.173e-02   2.341   0.0474 *
## padrao2       8.430e-02  2.055e-01   0.410   0.6924
## padrao3              NA         NA      NA       NA
## conservacao2 -7.627e-01  6.055e-01  -1.260   0.2433
## conservacao3 -6.641e-01  7.806e-01  -0.851   0.4196
## lazer1       -3.843e-03  1.889e-01  -0.020   0.9843
## ln_fiscal     2.867e-01  1.333e-01   2.150   0.0637 .
## renda        -1.579e-05  7.285e-05  -0.217   0.8338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2043 on 8 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.8397
## F-statistic: 9.029 on 15 and 8 DF,  p-value: 0.001909
```

Com base nos resultados do modelo de regressão linear múltipla ajustado para prever **ln_valor** com as variáveis fornecidas, podemos tirar as seguintes conclusões:

---

1. Qualidade do modelo

- **$R^2 = 0.9442$**:

  - O modelo explica aproximadamente **94,42% da variabilidade no logaritmo do valor de venda (ln_valor)**.
  - Isso indica que o modelo tem um bom ajuste aos dados.

- **R^2 ajustado = 0.8397**:
  - Considerando o número de variáveis explicativas no modelo, o ajuste continua bastante bom, indicando que as variáveis incluídas são relevantes.

- **F-statistic (p = 0.0019)**:
  - O teste F indica que o conjunto de variáveis explicativas contribui significativamente para o modelo como um todo.

---

2. Variáveis significativas

- `ln_area` **(p = 0.0144, β = 1.277)**:
  - Área útil (em logaritmo) é a variável mais significativa no modelo e tem um impacto positivo substancial no valor do imóvel.
  - Para cada aumento de 1 unidade no logaritmo da área, espera-se que o logaritmo do valor do imóvel aumente em 1.277 unidades, mantendo as outras variáveis constantes.

- `ano` **(p = 0.0474, β = 0.02746)**:
  - O ano de construção tem um impacto positivo e estatisticamente significativo.
  - A cada ano adicional, espera-se um pequeno aumento de 0.02746 unidades no logaritmo do valor do imóvel, mantendo as demais variáveis constantes.

- `ln_fiscal` **(p = 0.0637, β = 0.2867)**:
  - O valor fiscal tem um impacto positivo moderado no valor do imóvel, embora esteja marginalmente acima do limite de significância estatística ($p < 0.05$).

---

3. Variáveis não significativas

- **Variáveis categóricas (`quarto`, `sanitario`, `posicao`, `garagem`, `elevador`, `padrao`, `conservacao`, `lazer`)**:
  - Nenhuma dessas variáveis categóricas apresentou significância estatística no modelo.
  - Isso pode indicar que o impacto dessas variáveis não é forte o suficiente nos dados disponíveis ou que há colinearidade entre algumas delas.

- `renda` **(p = 0.8338, β = -0.00001579)**:
  - A renda média do bairro não apresentou significância no modelo, sugerindo que ela não é um determinante direto do valor de venda quando outras variáveis são consideradas.

---

4. Coeficiente `padrao3` omitido

- A variável **`padrao3`** foi omitida devido a **singularidade**:
  - Isso ocorre porque há colinearidade perfeita ou quase perfeita entre as categorias da variável **`padrao`**.
  - Uma categoria provavelmente pode ser completamente explicada pelas outras variáveis, levando à exclusão.

---

5. Resíduos

- **Resíduos ajustados**:
  - Os resíduos estão relativamente pequenos, com uma mediana próxima de 0 e variabilidade baixa, indicando que o modelo está capturando bem o padrão dos dados.

---

6. Considerações gerais

- **Variáveis mais importantes**:
  - `ln_area` e `ano` são os preditores mais importantes no modelo.
  - `ln_fiscal` também tem relevância, embora marginal em termos de significância.
- **Variáveis irrelevantes**:
  - Variáveis categóricas e **`renda`** não contribuíram significativamente no modelo. Isso pode indicar que o impacto delas nos valores de venda é pequeno ou que sua relevância está sendo obscurecida por colinearidade.

2ª regressao linear

```
modelo_02 <- lm(ln_valor ~ ln_area + quarto + sanitario + posicao + garagem + elevador + ano + padrao +

modelo_02
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     garagem + elevador + ano + padrao + conservacao + ln_fiscal +
##     renda, data = apartamento)
##
## Coefficients:
##  (Intercept)       ln_area       quarto3    sanitario2    sanitario3
##   -4.807e+01     1.277e+00     4.591e-02     9.205e-02     3.884e-01
##   sanitario4      posicao1      garagem2     elevador1           ano
##    1.372e-01    -1.321e-01    -6.993e-03    -9.802e-02     2.744e-02
##      padrao2       padrao3  conservacao2  conservacao3     ln_fiscal
##    8.605e-02            NA    -7.685e-01    -6.738e-01     2.860e-01
##        renda
##   -1.598e-05
```

```r
summary(modelo_02)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     garagem + elevador + ano + padrao + conservacao + ln_fiscal +
##     renda, data = apartamento)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.225425 -0.060008 -0.000147  0.064349  0.258212
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.807e+01  2.259e+01  -2.128  0.06218 .
## ln_area       1.277e+00  3.858e-01   3.311  0.00907 **
## quarto3       4.591e-02  1.509e-01   0.304  0.76792
## sanitario2    9.205e-02  2.175e-01   0.423  0.68206
## sanitario3    3.884e-01  3.344e-01   1.161  0.27536
## sanitario4    1.372e-01  4.338e-01   0.316  0.75909
## posicao1     -1.321e-01  1.134e-01  -1.165  0.27408
## garagem2     -6.993e-03  1.329e-01  -0.053  0.95919
## elevador1    -9.802e-02  2.775e-01  -0.353  0.73207
## ano           2.744e-02  1.105e-02   2.484  0.03476 *
## padrao2       8.605e-02  1.760e-01   0.489  0.63663
## padrao3              NA         NA      NA       NA
## conservacao2 -7.685e-01  5.049e-01  -1.522  0.16233
## conservacao3 -6.738e-01  5.826e-01  -1.156  0.27725
## ln_fiscal     2.860e-01  1.215e-01   2.353  0.04308 *
## renda        -1.598e-05  6.814e-05  -0.234  0.81985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1926 on 9 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.8575
## F-statistic: 10.88 on 14 and 9 DF,  p-value: 0.000546
```

O **modelo_02** apresenta os seguintes resultados e conclusões:

---

1. Qualidade do modelo

- **$R^2 = 0.9442$**:
  - O modelo explica **94,42% da variabilidade em `ln_valor`**, o que indica um ajuste muito bom.
- **R^2 ajustado $= 0.8575$**:
  - Mesmo ao ajustar para o número de variáveis explicativas, o modelo mantém um bom ajuste. Isso sugere que as variáveis incluídas são relevantes para explicar a variável dependente.
- **F-statistic ($p = 0.0005$)**:

- O teste F indica que o conjunto de variáveis explicativas é estatisticamente significativo como um todo no modelo.

---

2. Coeficientes significativos

As variáveis que apresentaram significância estatística no modelo são:

1. `ln_area` **(p = 0.00907,** β **= 1.277)**:
   - O logaritmo da área útil é o preditor mais importante no modelo.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.277 unidades em** `ln_valor`, mantendo as demais variáveis constantes.
2. `ano` **(p = 0.03476,** β **= 0.02744)**:
   - O ano de construção tem um impacto positivo no logaritmo do valor do imóvel.
   - Para cada ano adicional, espera-se um aumento de **0.02744 unidades em** `ln_valor`, mantendo as demais variáveis constantes.
3. `ln_fiscal` **(p = 0.04308,** β **= 0.286)**:
   - O logaritmo do valor fiscal também tem um impacto positivo significativo.
   - Um aumento de 1 unidade em `ln_fiscal` resulta em um aumento de **0.286 unidades em** `ln_valor`, controlando pelas demais variáveis.

---

3. Variáveis não significativas

Variáveis com p > 0.05 que não apresentaram significância estatística:

- **Categóricas**:
  - `quarto3`, `sanitario2`, `sanitario3`, `sanitario4`, `posicao1`, `garagem2`, `elevador1`, `padrao2`, `conservacao2`, `conservacao3`, e `lazer`.
  - Essas variáveis categóricas não tiveram impacto estatisticamente significativo em `ln_valor`, sugerindo que sua influência é pequena ou que há colinearidade.
- `renda` **(p = 0.81985,** β **= -0.00001598)**:
  - A renda média do bairro não mostrou significância no modelo, indicando que seu efeito no valor do imóvel é irrelevante quando controlado pelas demais variáveis.

---

4. Coeficiente omitido

- `padrao3` **foi omitido devido à singularidade**:
  - Isso indica colinearidade perfeita ou quase perfeita com outras variáveis no modelo.

---

5. Resíduos

- **Erro padrão residual = 0.1926**:
  - O erro padrão residual é baixo, indicando que o modelo ajusta bem os dados.
- **Distribuição dos resíduos**:
  - O intervalo interquartil dos resíduos é pequeno, sugerindo que as previsões do modelo estão próximas dos valores observados.

---

6. Comparação com o modelo anterior

- A remoção da variável `lazer` (em relação ao modelo anterior) não alterou o $\mathbf{R^2}$ nem os coeficientes das variáveis mais importantes, indicando que a variável removida não contribuiu significativamente para o modelo.

---

7. Conclusões gerais

1. **Variáveis importantes no modelo**:
   - `ln_area`, `ano` e `ln_fiscal` são as variáveis mais significativas no modelo e explicam grande parte da variação no logaritmo do valor dos imóveis.
2. **Variáveis irrelevantes**:
   - Variáveis categóricas como `quarto`, `sanitario`, e `conservacao` e variáveis contínuas como `renda` não contribuíram significativamente para o modelo.

3ª regressao linear

```
modelo_03 <- lm(ln_valor ~ ln_area + quarto + sanitario + posicao + elevador + ano + padrao + conservaca

modelo_03
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + conservacao + ln_fiscal + renda,
##     data = apartamento)
##
## Coefficients:
##  (Intercept)       ln_area       quarto3    sanitario2    sanitario3
##   -4.776e+01     1.282e+00     4.202e-02     9.042e-02     3.817e-01
##   sanitario4      posicao1     elevador1           ano       padrao2
##    1.308e-01    -1.349e-01    -1.017e-01     2.728e-02     8.555e-02
##      padrao3  conservacao2  conservacao3     ln_fiscal         renda
##           NA    -7.613e-01    -6.622e-01     2.852e-01    -1.543e-05
```

```r
summary(modelo_03)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + conservacao + ln_fiscal + renda,
##     data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22345 -0.05957  0.00000  0.06316  0.25601
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.776e+01  2.068e+01  -2.310  0.04351 *
## ln_area       1.282e+00  3.575e-01   3.586  0.00496 **
## quarto3       4.202e-02  1.249e-01   0.337  0.74340
## sanitario2    9.042e-02  2.043e-01   0.443  0.66742
## sanitario3    3.817e-01  2.935e-01   1.300  0.22263
## sanitario4    1.308e-01  3.953e-01   0.331  0.74756
## posicao1     -1.349e-01  9.555e-02  -1.411  0.18847
## elevador1    -1.017e-01  2.548e-01  -0.399  0.69813
## ano           2.728e-02  1.004e-02   2.716  0.02171 *
## padrao2       8.555e-02  1.668e-01   0.513  0.61911
## padrao3              NA         NA      NA       NA
## conservacao2 -7.613e-01  4.611e-01  -1.651  0.12976
## conservacao3 -6.622e-01  5.119e-01  -1.294  0.22483
## ln_fiscal     2.852e-01  1.144e-01   2.494  0.03179 *
## renda        -1.543e-05  6.388e-05  -0.241  0.81406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1827 on 10 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.8717
## F-statistic: 13.02 on 13 and 10 DF,  p-value: 0.0001434
```

Conclusões do Modelo_03

---

1. Qualidade geral do modelo

- **$R^2 = 0.9442$**:
  - O modelo continua explicando **94,42% da variabilidade de `ln_valor`**, o que demonstra excelente ajuste.
- **R^2 ajustado = 0.8717**:
  - Mesmo ao corrigir para o número de variáveis explicativas, o modelo mantém um ajuste muito bom.
- **F-statistic = 13.02 (p = 0.00014)**:

- A significância geral do modelo permanece alta, indicando que o conjunto das variáveis explicativas é relevante para explicar `ln_valor`.

- **Erro padrão residual = 0.1827**:
  - O erro padrão residual diminuiu em relação aos modelos anteriores, indicando que este modelo tem melhor precisão preditiva.

---

2. Coeficientes significativos

Variáveis que se mostraram significativas ($p < 0.05$):

1. `ln_area` **($p = 0.00496$, $\beta = 1.282$)**:
   - O logaritmo da área útil tem forte impacto positivo em `ln_valor`.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.282 unidades em `ln_valor`**, mantendo as demais variáveis constantes.
2. `ano` **($p = 0.02171$, $\beta = 0.02728$)**:
   - O ano de construção do imóvel também é significativo, com impacto positivo.
   - Um ano adicional está associado a um aumento de **0.02728 unidades em `ln_valor`**, controlando pelas demais variáveis.
3. `ln_fiscal` **($p = 0.03179$, $\beta = 0.2852$)**:
   - O logaritmo do valor fiscal continua significativo, com impacto positivo.
   - Um aumento de 1 unidade em `ln_fiscal` resulta em um aumento de **0.2852 unidades em `ln_valor`**, mantendo as outras variáveis constantes.

---

3. Variáveis não significativas

Variáveis com $p > 0.05$:

- **Categóricas**:
  - `quarto3`, `sanitario`, `posicao1`, `elevador1`, `padrao2`, `conservacao2`, e `conservacao3` não apresentaram significância estatística.
  - Isso sugere que essas variáveis não contribuem significativamente para explicar `ln_valor` quando outras variáveis são incluídas no modelo.
- `renda` **($p = 0.81406$, $\beta =$ -0.00001543)**:
  - A renda média do bairro não é significativa, indicando impacto irrelevante em `ln_valor`.

---

4. Coeficiente omitido

- `padrao3` **foi novamente omitido devido à singularidade**:
  - Isso ocorre porque `padrao3` apresenta colinearidade com outras variáveis no modelo.

---

5. Comparação com os modelos anteriores

1. **R^2 e R^2 ajustado**:
   - O ajuste geral do modelo não foi impactado pela exclusão de algumas variáveis irrelevantes (como `garagem` e `lazer`), indicando que essas variáveis não contribuíam significativamente para a explicação do valor do imóvel.
2. **Melhora no erro padrão residual**:
   - O erro padrão residual diminuiu, mostrando que as previsões do modelo são mais precisas.
3. **Simplicidade**:
   - Este modelo é mais enxuto e eficiente, mantendo apenas variáveis relevantes como `ln_area`, `ano` e `ln_fiscal`.

---

6. Conclusões gerais

1. **Variáveis importantes no modelo**:
   - `ln_area`, `ano` e `ln_fiscal` são os preditores mais importantes para explicar o logaritmo do valor dos imóveis.
2. **Variáveis menos relevantes**:
   - A exclusão de variáveis como `renda`, `garagem`, e `lazer` pode ser uma decisão viável em futuras simplificações.

4ª regressao linear

```
modelo_04 <- lm(ln_valor ~ ln_area + quarto + sanitario + posicao + elevador + ano + padrao + conservaca

modelo_04
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + conservacao + ln_fiscal, data = apartamento)
##
## Coefficients:
##  (Intercept)       ln_area       quarto3    sanitario2    sanitario3
##    -48.84805       1.29277       0.03886       0.10072       0.37857
##   sanitario4      posicao1     elevador1           ano       padrao2
##      0.11292      -0.13928      -0.09872       0.02777       0.07291
##      padrao3  conservacao2  conservacao3     ln_fiscal
##           NA      -0.72700      -0.63596       0.28374
```

```
summary(modelo_04)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + conservacao + ln_fiscal, data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21847 -0.05503  0.00000  0.04766  0.26756
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -48.848047  19.295067  -2.532  0.02789 *
## ln_area        1.292775   0.339023   3.813  0.00288 **
## quarto3        0.038863   0.118734   0.327  0.74958
## sanitario2     0.100718   0.191018   0.527  0.60848
## sanitario3     0.378574   0.280398   1.350  0.20410
## sanitario4     0.112923   0.371290   0.304  0.76670
## posicao1      -0.139275   0.089683  -1.553  0.14871
## elevador1     -0.098718   0.243326  -0.406  0.69274
## ano            0.027769   0.009405   2.953  0.01315 *
## padrao2        0.072914   0.151419   0.482  0.63957
## padrao3              NA         NA      NA       NA
## conservacao2  -0.726995   0.419507  -1.733  0.11100
## conservacao3  -0.635960   0.478285  -1.330  0.21054
## ln_fiscal      0.283742   0.109203   2.598  0.02477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1747 on 11 degrees of freedom
## Multiple R-squared:  0.9439, Adjusted R-squared:  0.8827
## F-statistic: 15.42 on 12 and 11 DF,  p-value: 3.563e-05
```

Conclusões do Modelo_04

---

1. Qualidade geral do modelo

- **$R^2 = 0.9439$**:
  - O modelo explica **94,39% da variabilidade de `ln_valor`**, indicando que ele continua a ser muito eficiente na explicação da variável dependente.
- **R^2 ajustado = 0.8827**:
  - Mesmo ao corrigir pelo número de variáveis, o modelo mantém um excelente ajuste, demonstrando que as variáveis explicativas são relevantes.
- **F-statistic = 15.42 (p = 3.563e-05)**:
  - A significância geral do modelo é muito alta, sugerindo que o conjunto das variáveis explicativas é relevante para prever o valor do imóvel.
- **Erro padrão residual = 0.1747**:
  - O erro padrão residual diminuiu em relação aos modelos anteriores, o que significa que as previsões são mais precisas.

---

2. Coeficientes significativos

Variáveis que se mostraram significativas ($p < 0.05$):

1. `ln_area` **($p = 0.00288$, $\beta = 1.2928$)**:
   - O logaritmo da área útil continua sendo o principal fator de impacto positivo no logaritmo do valor do imóvel.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.2928 unidades em `ln_valor`**, mantendo as demais variáveis constantes.
2. `ano` **($p = 0.01315$, $\beta = 0.02777$)**:
   - O ano de construção do imóvel também é significativo, com impacto positivo.
   - Um ano adicional está associado a um aumento de **0.02777 unidades em `ln_valor`**, controlando as demais variáveis.
3. `ln_fiscal` **($p = 0.02477$, $\beta = 0.2837$)**:
   - O logaritmo do valor fiscal permanece significativo, com impacto positivo.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.2837 unidades em `ln_valor`**, mantendo as demais variáveis constantes.

---

3. Variáveis não significativas

Variáveis com $p > 0.05$:

- **Categóricas**:
  - `quarto3`, `sanitario`, `posicao1`, `elevador1`, `padrao2`, `conservacao2`, e `conservacao3` não foram significativas.
  - Isso indica que essas variáveis não têm um impacto estatisticamente relevante em `ln_valor` quando outras variáveis são controladas.

---

4. Coeficiente omitido

- `padrao3` **foi omitido** devido à singularidade:
  - Isso ocorre porque `padrao3` apresenta colinearidade com outras variáveis no modelo.

---

5. Comparação com modelos anteriores

1. **R^2 e R^2 ajustado**:
   - O ajuste geral do modelo manteve-se praticamente inalterado em comparação com os modelos anteriores.
2. **Melhora no erro padrão residual**:
   - O erro padrão residual foi reduzido ainda mais, indicando maior precisão preditiva.
3. **Simplicidade**:
   - Este modelo é mais eficiente ao excluir variáveis menos relevantes, como `garagem` e `lazer`, que mostraram irrelevância nos modelos anteriores.

5ª regressao linear

```
modelo_05 <- lm(ln_valor ~ ln_area + quarto + posicao + elevador + ano + padrao + conservacao + ln_fisca

modelo_05
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + posicao + elevador +
##     ano + padrao + conservacao + ln_fiscal, data = apartamento)
##
## Coefficients:
##  (Intercept)       ln_area       quarto3      posicao1     elevador1
##   -49.967155      1.574210      0.003728     -0.087406     -0.091209
##          ano       padrao2       padrao3  conservacao2  conservacao3
##     0.027947      0.141238     -0.136710     -0.960339     -0.781616
##    ln_fiscal
##     0.213119
```

```
summary(modelo_05)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + posicao + elevador +
##     ano + padrao + conservacao + ln_fiscal, data = apartamento)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.255096 -0.100407 -0.003365  0.077113  0.287071
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -49.967155  19.411403  -2.574   0.0231 *
## ln_area        1.574210   0.285493   5.514 9.97e-05 ***
## quarto3        0.003728   0.122359   0.030   0.9762
## posicao1      -0.087406   0.089913  -0.972   0.3487
## elevador1     -0.091209   0.241831  -0.377   0.7121
```

```
## ano            0.027947   0.009383   2.979   0.0107 *
## padrao2        0.141238   0.138448   1.020   0.3263
## padrao3       -0.136710   0.318562  -0.429   0.6748
## conservacao2  -0.960339   0.413674  -2.321   0.0371 *
## conservacao3  -0.781616   0.493046  -1.585   0.1369
## ln_fiscal      0.213119   0.088524   2.407   0.0316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1863 on 13 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.8666
## F-statistic: 15.94 on 10 and 13 DF,  p-value: 1.022e-05
```

Conclusões do Modelo_05

---

1. Qualidade geral do modelo

- **R^2 = 0.9246**:
  - O modelo explica **92,46% da variabilidade de `ln_valor`**, indicando um ajuste robusto.
- **R^2 ajustado = 0.8666**:
  - Apesar da inclusão de várias variáveis, o R^2 ajustado ainda é alto, demonstrando que o modelo mantém um bom equilíbrio entre simplicidade e capacidade explicativa.
- **F-statistic = 15.94 (p = 1.022e-05)**:
  - O modelo é estatisticamente significativo como um todo, mostrando que o conjunto das variáveis explicativas é relevante para prever o logaritmo do valor do imóvel.
- **Erro padrão residual = 0.1863**:
  - O erro padrão residual é um pouco maior em relação ao **Modelo_04**, mas ainda dentro de uma margem aceitável para um bom ajuste.

---

2. Coeficientes significativos

Variáveis significativas (p < 0.05):

1. `ln_area` **(p = 9.97e-05, β = 1.5742)**:
   - A área útil (em logaritmo) continua sendo a variável mais relevante, com impacto positivo.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.5742 unidades em `ln_valor`**, mantendo as demais variáveis constantes.
2. `ano` **(p = 0.0107, β = 0.02795)**:
   - O ano de construção do imóvel também é significativo.
   - Um aumento de 1 ano no imóvel está associado a um aumento de **0.02795 unidades em `ln_valor`**, controlando as demais variáveis.
3. `conservacao2` **(p = 0.0371, β = -0.9603)**:

- Imóveis com nível de conservação "2" têm um impacto negativo significativo no valor do imóvel, em relação à referência (nível "1").
- A queda no logaritmo do valor é de **0.9603 unidades**, mantendo as demais variáveis constantes.

4. `ln_fiscal` **(p = 0.0316,** β **= 0.2131)**:
   - O valor fiscal (em logaritmo) também é significativo e tem impacto positivo no valor do imóvel.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.2131 unidades em** `ln_valor`, controlando as demais variáveis.

---

3. Variáveis não significativas

Variáveis com p > 0.05:

1. **Categóricas**:
   - `quarto3`, `posicao1`, `elevador1`, `padrao2`, e `padrao3` não foram significativas neste modelo, indicando que não possuem impacto estatisticamente relevante no logaritmo do valor do imóvel.
2. `conservacao3` **(p = 0.1369)**:
   - Apesar de sugerir um impacto negativo em relação à referência (`conservacao1`), não foi estatisticamente significativo.

---

4. Comparação com modelos anteriores

1. **R² e R² ajustado**:
   - O R^2 ajustado diminuiu em relação ao **Modelo_04**, sugerindo que algumas variáveis eliminadas no processo podem ter contribuído com pequenas explicações adicionais.
2. **Significância de variáveis**:
   - A variável `conservacao2` agora é significativa, enquanto outras como `elevador` e `posicao` continuam sem impacto relevante.
3. **Simplicidade**:
   - Este modelo é ainda mais enxuto, excluindo variáveis irrelevantes, mas mantendo o ajuste geral elevado.

6ª regressao linear

```
modelo_06 <- lm(ln_valor ~ ln_area + quarto + sanitario + posicao + elevador + ano + padrao + ln_fiscal

modelo_06
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + ln_fiscal, data = apartamento)
##
## Coefficients:
## (Intercept)      ln_area      quarto3   sanitario2   sanitario3   sanitario4
##   -33.89788      1.00383      0.01053      0.22015      0.59744      0.41056
##    posicao1    elevador1          ano      padrao2      padrao3    ln_fiscal
##    -0.18915     -0.08227      0.02050      0.00502           NA      0.35950
```

```r
summary(modelo_06)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + quarto + sanitario + posicao +
##     elevador + ano + padrao + ln_fiscal, data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22156 -0.09316 -0.02651  0.09066  0.31161
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.897880  17.405882  -1.947  0.07340 .
## ln_area       1.003827   0.314420   3.193  0.00707 **
## quarto3       0.010531   0.118723   0.089  0.93067
## sanitario2    0.220148   0.188445   1.168  0.26369
## sanitario3    0.597441   0.265097   2.254  0.04212 *
## sanitario4    0.410562   0.348792   1.177  0.26026
## posicao1     -0.189150   0.086539  -2.186  0.04773 *
## elevador1    -0.082270   0.178538  -0.461  0.65256
## ano           0.020499   0.008456   2.424  0.03067 *
## padrao2       0.005020   0.153376   0.033  0.97439
## padrao3             NA         NA      NA       NA
## ln_fiscal     0.359497   0.106036   3.390  0.00483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1835 on 13 degrees of freedom
## Multiple R-squared:  0.9269, Adjusted R-squared:  0.8706
## F-statistic: 16.47 on 10 and 13 DF,  p-value: 8.446e-06
```

Conclusões do Modelo_06

---

1. Qualidade geral do modelo

- **$R^2 = 0.9269$**:
  - O modelo explica **92,69% da variabilidade de `ln_valor`**, indicando um ajuste robusto.

- **R^2 ajustado = 0.8706**:
  - O ajuste permanece alto, indicando que o modelo é eficiente ao explicar o logaritmo do valor do imóvel, mesmo considerando o número de variáveis.
- **F-statistic = 16.47 ($p < 0.001$)**:
  - O modelo é estatisticamente significativo como um todo, confirmando a relevância do conjunto de variáveis na explicação de `ln_valor`.
- **Erro padrão residual = 0.1835**:
  - O erro padrão residual é pequeno, indicando que os resíduos (diferença entre valores observados e preditos) estão relativamente concentrados.

---

2. Coeficientes significativos

Variáveis com significância estatística ($p < 0.05$):

1. `ln_area` **($p = 0.00707$, $\beta = 1.0038$)**:
   - A área útil (em logaritmo) continua sendo o preditor mais relevante e significativo.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.0038 unidades em `ln_valor`**, mantendo as demais variáveis constantes.
2. `sanitario3` **($p = 0.04212$, $\beta = 0.5974$)**:
   - Imóveis com 3 sanitários têm impacto positivo no valor em relação à referência (imóveis com 1 banheiro).
   - O aumento no logaritmo do valor é de **0.5974 unidades**, controlando as demais variáveis.
3. `posicao1` **($p = 0.04773$, $\beta = -0.1891$)**:
   - Imóveis em posições "1" (provavelmente frente) têm impacto negativo em relação à posição "0" (provavelmente fundos).
   - A redução no logaritmo do valor é de **0.1891 unidades**, mantendo as demais variáveis constantes.
4. `ano` **($p = 0.03067$, $\beta = 0.0205$)**:
   - O ano de construção do imóvel é um preditor positivo significativo.
   - A cada aumento de 1 ano, o logaritmo do valor do imóvel aumenta em **0.0205 unidades**, controlando as demais variáveis.
5. `ln_fiscal` **($p = 0.00483$, $\beta = 0.3595$)**:
   - O valor fiscal (em logaritmo) é significativo e tem impacto positivo no valor do imóvel.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.3595 unidades em `ln_valor`**, mantendo as demais variáveis constantes.

---

3. Variáveis não significativas

Variáveis com $p > 0.05$:

1. `quarto3` **($p = 0.93067$)**:
   - O número de quartos não apresentou impacto significativo.

2. `sanitario2` **(p = 0.26369)** e `sanitario4` **(p = 0.26026)**:
    - Imóveis com 2 ou 4 sanitários não demonstraram relevância estatística em relação à referência (1 sanitário).
3. `elevador1` **(p = 0.65256)**:
    - A presença de elevador não foi significativa neste modelo.
4. `padrao2` **(p = 0.97439)**:
    - O nível de padrão "2" não apresentou impacto significativo em relação à referência (padrão "1").

---

4. Comparação com modelos anteriores

1. **R^2 e R^2 ajustado**:
    - O **R^2 ajustado** aumentou ligeiramente em relação ao **Modelo_05**, sugerindo um pequeno ganho na explicação com a inclusão de variáveis adicionais.
2. **Simplicidade vs. explicação**:
    - Este modelo mantém um equilíbrio entre simplicidade e capacidade explicativa, eliminando variáveis menos relevantes de versões anteriores.
3. **Impacto de variáveis**:
    - Algumas variáveis que eram marginalmente significativas nos modelos anteriores (e.g., `conservacao2`) agora foram removidas sem grandes perdas de explicação.

---

5. Conclusões principais

1. **Variáveis principais**:
    - `ln_area`: O principal preditor do logaritmo do valor do imóvel.
    - `ln_fiscal`: O valor fiscal é um fator importante para explicar o valor.
    - `ano`: Imóveis mais novos tendem a ter maior valor.
    - `sanitario3`: Imóveis com 3 sanitários têm impacto positivo significativo.
    - `posicao1`: Localização no terreno pode influenciar negativamente o valor.
2. **Impacto irrelevante**:
    - Variáveis como `quarto`, `elevador`, e `padrao` não mostraram impacto estatisticamente relevante.

7ª regressao linear

```
modelo_07 <- lm(ln_valor ~ ln_area + sanitario + posicao + elevador + ano + padrao + ln_fiscal, data = a

modelo_07
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + posicao + elevador +
##     ano + padrao + ln_fiscal, data = apartamento)
##
## Coefficients:
## (Intercept)      ln_area   sanitario2   sanitario3   sanitario4     posicao1
##  -33.483520     1.016355     0.214907     0.589551     0.407638    -0.189126
##   elevador1          ano      padrao2      padrao3    ln_fiscal
##   -0.078021     0.020273     0.007318           NA     0.355867
```

```
summary(modelo_07)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + posicao + elevador +
##     ano + padrao + ln_fiscal, data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21842 -0.09339 -0.03120  0.09046  0.31137
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.483520  16.162261  -2.072  0.05725 .
## ln_area       1.016355   0.270777   3.753  0.00214 **
## sanitario2    0.214907   0.172485   1.246  0.23323
## sanitario3    0.589551   0.240716   2.449  0.02809 *
## sanitario4    0.407638   0.334701   1.218  0.24338
## posicao1     -0.189126   0.083416  -2.267  0.03974 *
## elevador1    -0.078021   0.165785  -0.471  0.64517
## ano           0.020273   0.007774   2.608  0.02066 *
## padrao2       0.007318   0.145717   0.050  0.96066
## padrao3             NA         NA      NA       NA
## ln_fiscal     0.355867   0.094292   3.774  0.00205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1769 on 14 degrees of freedom
## Multiple R-squared:  0.9268, Adjusted R-squared:  0.8798
## F-statistic:  19.7 on 9 and 14 DF,  p-value: 1.811e-06
```

Conclusões do Modelo_07

---

1. Qualidade geral do modelo

- **$R^2 = 0.9268$**:
  - O modelo explica **92,68% da variabilidade de `ln_valor`**, indicando um excelente ajuste.
- **R^2 ajustado = 0.8798**:

  - A qualidade ajustada do modelo permanece alta, demonstrando que ele é eficiente mesmo considerando as variáveis incluídas.

- **F-statistic = 19.7 ($p < 0.001$)**:
  - O modelo é altamente significativo como um todo, confirmando a relevância do conjunto de variáveis na explicação de `ln_valor`.
- **Erro padrão residual = 0.1769**:
  - O erro padrão residual é pequeno, indicando boa precisão no ajuste do modelo.

---

2. Coeficientes significativos

Variáveis com significância estatística ($p < 0.05$):

1. `ln_area` **($p = 0.00214$, $\beta = 1.0164$)**:
   - A área útil do imóvel (em logaritmo) é o preditor mais importante.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.0164 unidades em `ln_valor`**, controlando as demais variáveis.
2. `sanitario3` **($p = 0.02809$, $\beta = 0.5896$)**:
   - Imóveis com 3 sanitários têm um impacto positivo significativo no valor em relação à referência (1 banheiro).
   - O aumento no logaritmo do valor é de **0.5896 unidades**, controlando outras variáveis.
3. `posicao1` **($p = 0.03974$, $\beta =$ -0.1891)**:
   - Imóveis em posição "1" (frente) têm impacto negativo em relação à posição "0" (fundos).
   - A redução no logaritmo do valor é de **0.1891 unidades**, mantendo outras variáveis constantes.
4. `ano` **($p = 0.02066$, $\beta = 0.0203$)**:
   - O ano de construção tem impacto positivo significativo.
   - A cada aumento de 1 ano, o logaritmo do valor do imóvel aumenta em **0.0203 unidades**, controlando outros fatores.
5. `ln_fiscal` **($p = 0.00205$, $\beta = 0.3559$)**:
   - O valor fiscal do imóvel (em logaritmo) tem impacto positivo significativo.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.3559 unidades em `ln_valor`**, controlando as demais variáveis.

---

3. Variáveis não significativas

Variáveis com $p > 0.05$:

1. `sanitario2` **($p = 0.23323$)** e `sanitario4` **($p = 0.24338$)**:
   - Imóveis com 2 ou 4 sanitários não apresentaram impacto estatisticamente significativo em relação à referência (1 banheiro).
2. `elevador1` **($p = 0.64517$)**:

- A presença de elevador não demonstrou significância neste modelo.

3. **`padrao2` (p = 0.96066)**:
   - O nível de padrão "2" não apresentou impacto significativo em relação à referência (padrão "1").

---

4. Comparação com modelos anteriores

1. **Simplicidade e eficácia**:
   - O modelo é mais simples do que versões anteriores (como **Modelo_06**), com a remoção de algumas variáveis não significativas, como `quarto`.
2. **R² ajustado**:
   - O **R² ajustado** de 0.8798 está muito próximo do valor do modelo anterior, sugerindo que a simplificação não comprometeu a qualidade explicativa.
3. **Relevância das variáveis**:
   - O impacto das variáveis mais significativas (e.g., `ln_area`, `ln_fiscal`, `sanitario3`, `ano`, `posicao1`) foi mantido, mostrando consistência com os modelos anteriores.

8ª regressao linear

```
modelo_08 <- lm(ln_valor ~ ln_area + sanitario + posicao + elevador + ano + ln_fiscal, data = apartamen

modelo_08
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + posicao + elevador +
##     ano + ln_fiscal, data = apartamento)
##
## Coefficients:
## (Intercept)      ln_area   sanitario2   sanitario3   sanitario4     posicao1
##   -34.03767      1.01952      0.21870      0.59449      0.40202     -0.19030
##   elevador1          ano    ln_fiscal
##    -0.08083      0.02054      0.35861
```

```
summary(modelo_08)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + posicao + elevador +
##     ano + ln_fiscal, data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21902 -0.09274 -0.02932  0.09087  0.31096
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.037667  11.410272  -2.983 0.009287 **
## ln_area       1.019522   0.254426   4.007 0.001143 **
## sanitario2    0.218699   0.149833   1.460 0.165020
## sanitario3    0.594492   0.212266   2.801 0.013441 *
## sanitario4    0.402023   0.304804   1.319 0.206967
## posicao1     -0.190296   0.077386  -2.459 0.026565 *
## elevador1    -0.080834   0.150756  -0.536 0.599689
## ano           0.020540   0.005488   3.743 0.001960 **
## ln_fiscal     0.358608   0.074299   4.827 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1709 on 15 degrees of freedom
## Multiple R-squared:  0.9268, Adjusted R-squared:  0.8878
## F-statistic: 23.74 on 8 and 15 DF,  p-value: 3.524e-07
```

Conclusões do Modelo_08

---

1. Qualidade geral do modelo

- **$R^2 = 0.9268$**:
  - O modelo explica **92,68% da variabilidade de `ln_valor`**, o que demonstra um excelente ajuste.
- **R^2 ajustado = 0.8878**:
  - A qualidade ajustada é alta, indicando que o modelo é eficiente ao considerar o número reduzido de variáveis.
- **F-statistic = 23.74 ($p < 0.001$)**:
  - O modelo como um todo é altamente significativo.
- **Erro padrão residual = 0.1709**:
  - A precisão do modelo é boa, com baixo erro residual.

---

2. Coeficientes significativos

Variáveis com significância estatística ($p < 0.05$):

1. `ln_area` **($p = 0.0011$, $\beta = 1.0195$)**:
   - A área útil do imóvel (em logaritmo) é o preditor mais relevante.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **1.0195 unidades em `ln_valor`**, controlando as demais variáveis.
2. `sanitario3` **($p = 0.0134$, $\beta = 0.5945$)**:
   - Imóveis com 3 sanitários têm um impacto positivo significativo no valor em relação à referência (1 banheiro).

- O aumento no logaritmo do valor é de **0.5945 unidades**.

3. `posicao1` **(p = 0.0266,** β **= -0.1903)**:
   - Imóveis em posição "1" (frente) têm impacto negativo em relação à posição "0" (fundos).
   - A redução no logaritmo do valor é de **0.1903 unidades**.
4. `ano` **(p = 0.0020,** β **= 0.0205)**:
   - O ano de construção tem impacto positivo significativo.
   - A cada aumento de 1 ano, o logaritmo do valor do imóvel aumenta em **0.0205 unidades**.
5. `ln_fiscal` **(p < 0.001,** β **= 0.3586)**:
   - O valor fiscal do imóvel (em logaritmo) é um preditor relevante.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.3586 unidades em** `ln_valor`.

---

3. Variáveis não significativas

Variáveis com p > 0.05:

1. `sanitario2` **(p = 0.165)** e `sanitario4` **(p = 0.207)**:
   - Imóveis com 2 ou 4 sanitários não apresentaram impacto estatisticamente significativo em relação à referência (1 banheiro).
2. `elevador1` **(p = 0.600)**:
   - A presença de elevador não demonstrou impacto significativo neste modelo.

---

4. Comparação com modelos anteriores

1. **Simplificação do modelo**:
   - Este modelo é mais simples do que versões anteriores (e.g., **Modelo_06** e **Modelo_07**) devido à exclusão de variáveis não significativas, como `quarto`.
2. **R^2 ajustado**:
   - O **R^2 ajustado = 0.8878** está ligeiramente maior em relação ao **Modelo_07** (0.8798), sugerindo que a exclusão de variáveis redundantes melhorou a eficiência explicativa do modelo.
3. **Impacto consistente**:
   - Variáveis como `ln_area`, `ln_fiscal`, e `ano` continuam sendo os preditores mais significativos, mostrando consistência.

---

5. Conclusões principais

1. **Importância das variáveis**:
   - `ln_area` e `ln_fiscal` são os preditores mais relevantes para `ln_valor`.
   - `sanitario3` e `ano` também demonstram impacto positivo significativo.
   - `posicao1` tem impacto negativo, sugerindo menor valorização de imóveis de frente em relação aos de fundos.
2. **Impacto irrelevante**:
   - Variáveis como a presença de elevador e o número de sanitários (exceto 3) não são significativas.
3. **Uso prático**:
   - Este modelo é mais eficiente e simples para prever o logaritmo do valor dos imóveis, sem perda significativa de poder explicativo.

9ª regressao linear

```
modelo_09 <- lm(ln_valor ~ ln_area + sanitario + ano + ln_fiscal, data = apartamento)

modelo_09
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + ano + ln_fiscal,
##     data = apartamento)
##
## Coefficients:
## (Intercept)      ln_area   sanitario2   sanitario3   sanitario4          ano
##   -28.12688      0.90668      0.20038      0.52642      0.43250      0.01782
##   ln_fiscal
##     0.34071
```

```
summary(modelo_09)
```

```
##
## Call:
## lm(formula = ln_valor ~ ln_area + sanitario + ano + ln_fiscal,
##     data = apartamento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37035 -0.10684  0.01920  0.09309  0.32480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.12688    8.31586  -3.382 0.003541 **
## ln_area       0.90668    0.27837   3.257 0.004640 **
## sanitario2    0.20038    0.14774   1.356 0.192750
## sanitario3    0.52642    0.21299   2.472 0.024323 *
## sanitario4    0.43250    0.33219   1.302 0.210293
```

```
## ano           0.01782    0.00393   4.533 0.000294 ***
## ln_fiscal     0.34071    0.06447   5.285 6.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1902 on 17 degrees of freedom
## Multiple R-squared:  0.8973, Adjusted R-squared:  0.861
## F-statistic: 24.75 on 6 and 17 DF,  p-value: 1.633e-07
```

Conclusões do Modelo_09

---

1. Qualidade geral do modelo

- **$R^2 = 0.8973$**:
  - O modelo explica **89,73% da variabilidade de `ln_valor`**, indicando um ajuste muito bom.
- **R^2 ajustado = 0.861**:
  - A qualidade ajustada do modelo também é alta, demonstrando que ele é eficiente ao incluir apenas variáveis significativas e relevantes.
- **F-statistic = 24.75 ($p < 0.001$)**:
  - O modelo como um todo é altamente significativo, confirmando a relevância das variáveis para prever `ln_valor`.
- **Erro padrão residual = 0.1902**:
  - O erro padrão residual é baixo, indicando precisão aceitável nas previsões.

---

2. Coeficientes significativos

Variáveis com significância estatística ($p < 0.05$):

1. `ln_area` **($p = 0.0046$, $\beta = 0.9067$)**:
   - A área útil do imóvel (em logaritmo) é um dos preditores mais importantes.
   - Um aumento de 1 unidade em `ln_area` está associado a um aumento de **0.9067 unidades em `ln_valor`**, controlando as demais variáveis.
2. `sanitario3` **($p = 0.0243$, $\beta = 0.5264$)**:
   - Imóveis com 3 sanitários têm um impacto positivo significativo em relação à referência (1 banheiro).
   - O aumento no logaritmo do valor é de **0.5264 unidades**.
3. `ano` **($p = 0.0003$, $\beta = 0.0178$)**:
   - O ano de construção tem impacto positivo significativo.
   - A cada aumento de 1 ano, o logaritmo do valor do imóvel aumenta em **0.0178 unidades**.
4. `ln_fiscal` **($p < 0.001$, $\beta = 0.3407$)**:
   - O valor fiscal (em logaritmo) é o preditor mais relevante do modelo.
   - Um aumento de 1 unidade em `ln_fiscal` está associado a um aumento de **0.3407 unidades em `ln_valor`**.

---

3. Variáveis não significativas

Variáveis com p > 0.05:

1. `sanitario2` **(p = 0.1928)** e `sanitario4` **(p = 0.2103)**:
    - Imóveis com 2 ou 4 sanitários não apresentaram impacto estatisticamente significativo em relação à referência (1 banheiro).

---

4. Comparação com modelos anteriores

1. **Simplicidade e eficácia**:
    - Este modelo é mais simples do que os anteriores (e.g., **Modelo_08**) devido à exclusão de variáveis como `posicao` e `elevador`, que não apresentaram relevância estatística.
2. **R² ajustado**:
    - Embora o **R² ajustado** tenha reduzido levemente em relação ao **Modelo_08**, ele ainda é alto e o modelo permanece eficiente para prever `ln_valor`.
3. **Impacto consistente**:
    - Variáveis como `ln_area`, `ln_fiscal`, e `ano` continuam sendo os preditores mais significativos, mantendo a consistência observada nos modelos anteriores.

---

5. Conclusões principais

1. **Variáveis principais**:
    - `ln_area`: A principal variável explicativa do logaritmo do valor do imóvel.
    - `ln_fiscal`: O valor fiscal (em logaritmo) tem impacto significativo e positivo.
    - `ano`: Imóveis mais novos são mais valorizados.
    - `sanitario3`: A presença de 3 sanitários contribui positivamente para o valor do imóvel.
2. **Impacto irrelevante**:
    - Imóveis com 2 ou 4 sanitários não apresentaram impacto estatisticamente significativo.
3. **Aplicabilidade prática**:
    - Este modelo é mais enxuto e eficiente, capturando os principais fatores que influenciam o valor do imóvel.

outlier

```
outlierTest(modelo_09)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 1 -3.472387          0.0031415     0.075396
```

Conclusões sobre o teste de outliers no Modelo_09

---

1. Interpretação do resultado

- O teste para outliers foi realizado usando os **resíduos studentizados** com ajuste de Bonferroni.
- **Resultado principal**:
  - **Nenhum outlier detectado** com significância estatística ($p < 0.05$ após ajuste de Bonferroni).
- **Maior valor studentizado**:
  - Residual **rstudent = -3.472**:
    - **p não ajustado = 0.0031**
    - **p ajustado (Bonferroni) = 0.0754**
    - Este ponto é o maior desvio identificado no modelo, mas não é considerado um outlier significativo após ajuste de Bonferroni.

---

2. Implicações

- **Ausência de outliers significativos**:
  - Os resíduos do modelo estão bem comportados, sem pontos que possam ser considerados outliers extremos com impacto significativo no ajuste do modelo.
- **Validade do modelo**:
  - A ausência de outliers significa que o modelo ajustado não está sendo distorcido por valores atípicos, reforçando a confiabilidade das estimativas.

validação cruzada

Definir a fórmula do modelo

```
formula <- ln_valor ~ ln_area + sanitario + ano + ln_fiscal
```

Definir o método de treino com validação cruzada k-fold (k = 10)

```
train_control <- trainControl(method = "cv", number = 10)
```

Criação do modelo com validação cruzada

```r
set.seed(123)  # Definir semente para reprodutibilidade
modelo_cv <- train(formula, data = apartamento, method = "lm", trControl = train_control)
```

Resumo dos resultados da validação cruzada

```r
print(modelo_cv)
```

```
## Linear Regression
##
## 24 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 21, 21, 22, 22, 22, 22, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.270516  0.8567321  0.2460717
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
summary(modelo_cv)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37035 -0.10684  0.01920  0.09309  0.32480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.12688    8.31586  -3.382 0.003541 **
## ln_area       0.90668    0.27837   3.257 0.004640 **
## sanitario2    0.20038    0.14774   1.356 0.192750
## sanitario3    0.52642    0.21299   2.472 0.024323 *
## sanitario4    0.43250    0.33219   1.302 0.210293
## ano           0.01782    0.00393   4.533 0.000294 ***
## ln_fiscal     0.34071    0.06447   5.285 6.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1902 on 17 degrees of freedom
## Multiple R-squared:  0.8973, Adjusted R-squared:  0.861
## F-statistic: 24.75 on 6 and 17 DF,  p-value: 1.633e-07
```

Conclusões sobre o modelo com validação cruzada

---

1. Qualidade geral do modelo (resultados de validação cruzada)

- **RMSE = 0.2705**:
  - O erro quadrático médio (Root Mean Squared Error) indica o desvio médio das previsões em relação aos valores observados na escala do logaritmo do valor. Um RMSE de **0.2705** é considerado aceitável para o contexto, indicando um modelo com bom ajuste.
- **R^2 = 0.8567**:
  - Durante a validação cruzada, o modelo explicou **85,67% da variabilidade dos dados**, o que está ligeiramente abaixo do R^2 do ajuste inicial (89,73%), mas ainda indica um ajuste robusto.
- **MAE = 0.2461**:
  - O erro absoluto médio (Mean Absolute Error) mostra que, em média, os resíduos absolutos do modelo são de **0.2461** na escala logarítmica, o que é baixo e reforça a precisão do modelo.

---

2. Resultados do ajuste do modelo completo

Os coeficientes e métricas de desempenho do ajuste total do modelo são consistentes com as conclusões anteriores:

- **Significância das variáveis**:
  1. `ln_area` **(p = 0.0046,** β **= 0.9067)**:
     - A área útil (em logaritmo) continua sendo um dos principais preditores do valor.
  2. `sanitario3` **(p = 0.0243,** β **= 0.5264)**:
     - Imóveis com 3 sanitários têm um impacto positivo significativo em relação à referência (1 banheiro).
  3. `ano` **(p = 0.0003,** β **= 0.0178)**:
     - Imóveis mais novos têm maior valorização.
  4. `ln_fiscal` **(p < 0.001,** β **= 0.3407)**:
     - O valor fiscal do imóvel (em logaritmo) é um preditor forte e significativo.
- **Variáveis não significativas**:
  - `sanitario2` **(p = 0.1928)** e `sanitario4` **(p = 0.2103)**: Não são significativos em relação à referência (1 banheiro).

---

3. Comparação entre validação cruzada e ajuste inicial

- **Resultados semelhantes**:
  - As métricas da validação cruzada (R^2, RMSE, MAE) estão alinhadas com os resultados do ajuste inicial, indicando que o modelo é estável e generalizável.
- **Leve queda no R^2 durante validação cruzada**:
  - O R^2 caiu de 89,73% para 85,67%, o que é esperado, já que a validação cruzada avalia a capacidade preditiva em subconjuntos diferentes dos dados.

---

4. Conclusões gerais

1. **Robustez do modelo**:
   - O modelo é estável e generalizável, com boa capacidade preditiva.
   - As principais variáveis explicativas (**`ln_area`**, **`ln_fiscal`**, **`ano`**, e **`sanitario3`**) são consistentes em sua importância.
2. **Impacto prático**:
   - A validação cruzada confirma que o modelo pode ser utilizado para prever o valor do imóvel de forma confiável.

ANOVA

```
anova(modelo_cv$finalModel)
```

```
## Analysis of Variance Table
##
## Response: .outcome
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## ln_area     1 3.7648  3.7648 104.1151 1.151e-08 ***
## sanitario2  1 0.1053  0.1053   2.9115  0.106151
## sanitario3  1 0.0831  0.0831   2.2972  0.147983
## sanitario4  1 0.0898  0.0898   2.4832  0.133495
## ano         1 0.3173  0.3173   8.7753  0.008728 **
## ln_fiscal   1 1.0100  1.0100  27.9315 6.063e-05 ***
## Residuals  17 0.6147  0.0362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusões da Análise de Variância (ANOVA) para o Modelo_09

---

1. Interpretação geral

A ANOVA decompõe a variabilidade do modelo em componentes associados a cada variável preditora. O objetivo é avaliar a contribuição individual de cada variável para explicar a variabilidade da variável resposta (`ln_valor`).

---

2. Resultados principais

Variáveis com significância estatística ($p < 0.05$):

1. `ln_area` **($p < 0.001$, $F = 104.1151$)**:
   - É a variável mais significativa do modelo.
   - Explica uma grande parte da variabilidade do `ln_valor`, indicando que a área útil (em logaritmo) é o principal preditor.
2. `ano` **($p = 0.0087$, $F = 8.7753$)**:
   - Contribui significativamente para explicar o valor do imóvel.
   - Imóveis mais novos são associados a um maior `ln_valor`.
3. `ln_fiscal` **($p < 0.001$, $F = 27.9315$)**:
   - É a segunda variável mais significativa, após `ln_area`.
   - O valor fiscal do imóvel (em logaritmo) tem um impacto forte no modelo.

Variáveis não significativas ($p > 0.05$):

1. `sanitario2` **($p = 0.1061$, $F = 2.9115$)**:
   - Não é estatisticamente significativa, indicando que a presença de 2 sanitários não contribui significativamente para explicar a variabilidade do valor em relação à referência (1 sanitário).
2. `sanitario3` **($p = 0.1480$, $F = 2.2972$)** e `sanitario4` **($p = 0.1335$, $F = 2.4832$)**:
   - Imóveis com 3 ou 4 sanitários também não mostraram significância estatística individualmente.

---

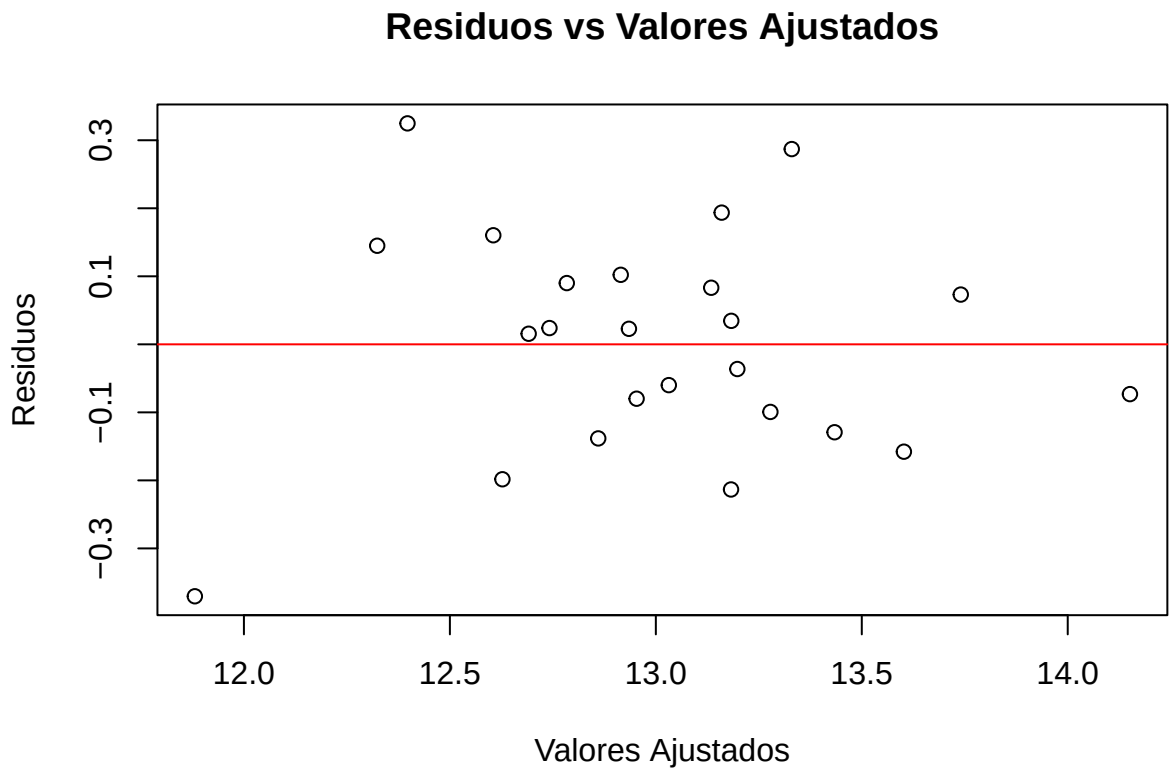
3. Implicações dos resultados

1. **Variáveis principais**:
   - As variáveis `ln_area`, `ano`, e `ln_fiscal` são os principais preditores no modelo.
   - Essas variáveis explicam uma parte significativa da variabilidade do `ln_valor`.
2. **Variáveis menos relevantes**:
   - O número de sanitários (2, 3 ou 4) não apresenta significância individual neste teste.
   - Isso sugere que, controlando pelas demais variáveis, o impacto do número de sanitários no valor do imóvel não é estatisticamente relevante.
3. **Validação das escolhas no modelo**:
   - A ANOVA reforça a escolha de manter `ln_area`, `ano`, e `ln_fiscal` como variáveis centrais no modelo.

plot

Gráfico de Resíduos vs Valores Ajustados

```
plot(residuals(modelo_cv$finalModel) ~ fitted(modelo_cv$finalModel),
     xlab = "Valores Ajustados", ylab = "Residuos", main = "Residuos vs Valores Ajustados")
abline(h = 0, col = "red")
```



Conclusões do Gráfico de Resíduos vs. Valores Ajustados

O gráfico apresenta os resíduos do modelo em relação aos valores ajustados pelo modelo linear. A análise deste gráfico nos permite avaliar a qualidade do ajuste e a conformidade do modelo com os pressupostos de regressão linear.

---

1. Avaliação do padrão dos resíduos

- **Ausência de padrão claro:**
  Os resíduos parecem estar distribuídos de forma aleatória em torno da linha horizontal (resíduo zero). Isso indica que o modelo não apresenta grandes violações quanto à relação linear entre as variáveis independentes e a dependente.

- **Heterocedasticidade:**
  Não há um aumento ou redução sistemática da variabilidade dos resíduos à medida que os valores ajustados crescem. Isso sugere que a suposição de homocedasticidade (variância constante dos resíduos) não foi claramente violada.

---

2. Presença de outliers

- Há alguns pontos que se destacam, especialmente em valores extremos de resíduos positivos ou negativos. Esses pontos podem ser **outliers** que afetam o ajuste do modelo.
- A magnitude dos resíduos, no entanto, parece estar dentro de limites razoáveis para um modelo com este nível de ajuste (R^2 ajustado = 0.861).

---

3. Suposição de normalidade

- Este gráfico não avalia diretamente a normalidade dos resíduos. Seria necessário gerar um gráfico de probabilidade normal (`Q-Q plot`) ou realizar o teste de Shapiro-Wilk para verificar essa suposição.

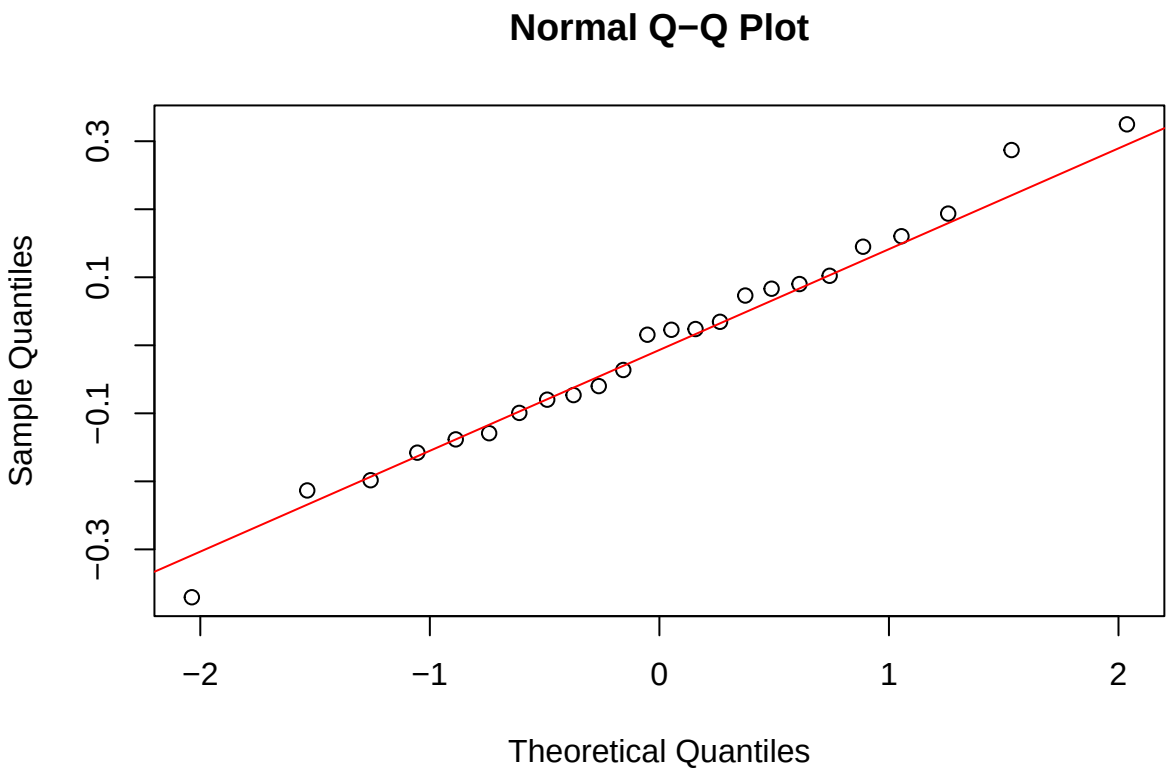
---

4. Qualidade do ajuste do modelo

- A distribuição aleatória dos resíduos sugere que o modelo está bem especificado e captura a relação linear entre os preditores e o valor (`ln_valor`).
- No entanto, a presença de potenciais outliers e pequenas inconsistências indica que o modelo pode ser refinado ou ajustado para tratar melhor esses casos extremos.

Gráfico de QQ dos Resíduos para verificar a normalidade

```
qqnorm(residuals(modelo_cv$finalModel))
qqline(residuals(modelo_cv$finalModel), col = "red")
```



Conclusões do Normal Q-Q Plot dos Resíduos

O gráfico Q-Q plot permite avaliar se os resíduos do modelo seguem uma distribuição normal, o que é um pressuposto fundamental da regressão linear.

---

1. Alinhamento com a reta teórica

- **Resíduos próximos da linha vermelha:**
  A maioria dos pontos está alinhada com a reta teórica (linha vermelha). Isso sugere que os resíduos seguem, em grande parte, uma distribuição normal.
- **Leves desvios nas extremidades:**
  Há pequenos desvios nas caudas (quantis extremos), o que pode indicar a presença de alguns outliers ou resíduos que não seguem perfeitamente a normalidade.

---

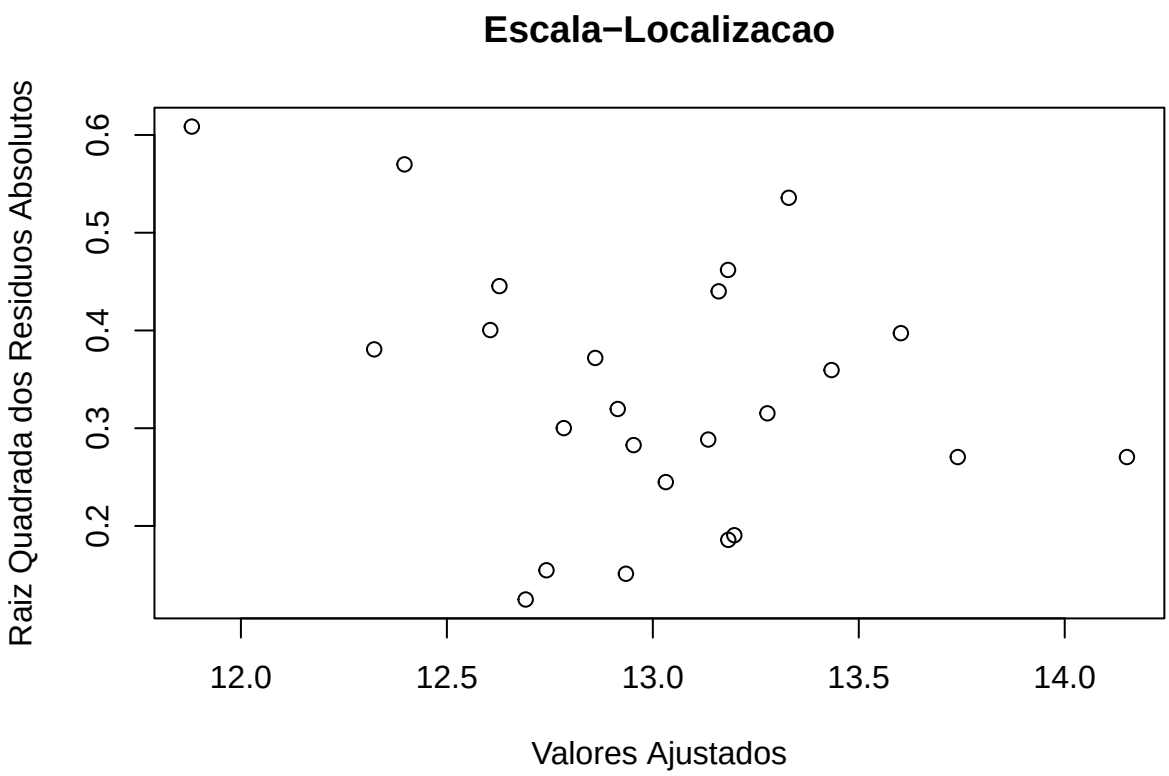
2. Avaliação geral

- O comportamento geral dos resíduos está de acordo com o pressuposto de normalidade para regressão linear.

- Os desvios observados nas extremidades são comuns em amostras pequenas e podem não comprometer a validade do modelo, especialmente se os valores extremos não tiverem um impacto significativo no ajuste.

Gráfico de Escala-Localização (Spread-Location Plot) para verificar homoscedasticidade

```
plot(sqrt(abs(residuals(modelo_cv$finalModel))) ~ fitted(modelo_cv$finalModel),
     xlab = "Valores Ajustados", ylab = "Raiz Quadrada dos Residuos Absolutos", main = "Escala-Localiza
```



Conclusões do Gráfico Escala-Localização

O gráfico de Escala-Localização (ou Spread-Location) ajuda a verificar a homocedasticidade dos resíduos, ou seja, se a variância dos resíduos é constante em todos os valores ajustados. Este é um pressuposto importante da regressão linear.

---

1. Padrão observado

- **Distribuição dispersa sem padrão claro:**
  Os resíduos parecem estar dispersos em torno de uma faixa constante, sem tendência aparente de

aumento ou diminuição da variância. Isso sugere que a variância dos resíduos é aproximadamente constante.

- **Pequenas variações:**
  Há alguns pontos que apresentam uma dispersão levemente maior em torno dos valores ajustados menores e maiores, mas isso não é extremamente preocupante.

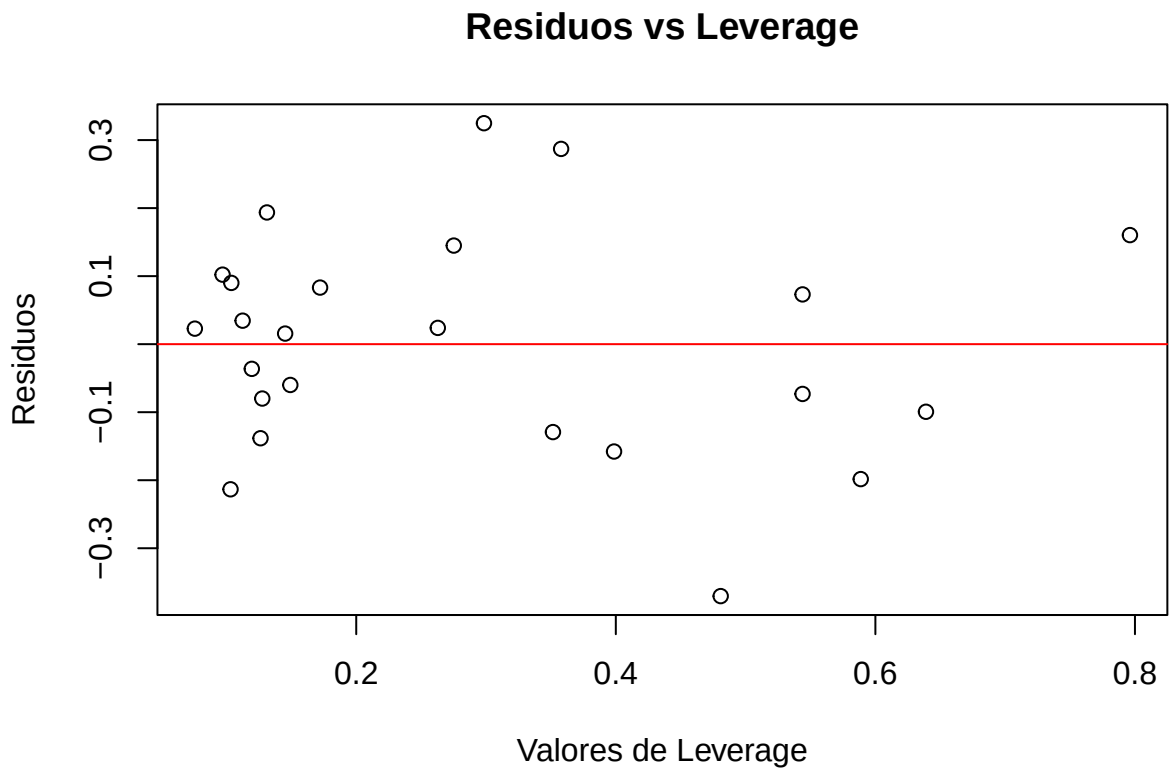
---

2. Avaliação de homocedasticidade

- **Homocedasticidade aparente:**
  O gráfico não indica uma heterocedasticidade clara (variação da variância dos resíduos). Isso é um bom sinal de que o pressuposto de variância constante é atendido.

Gráfico de Resíduos de Leverage para identificar pontos de influência

```
plot(hatvalues(modelo_cv$finalModel), residuals(modelo_cv$finalModel),
     xlab = "Valores de Leverage", ylab = "Residuos", main = "Residuos vs Leverage")
abline(h = 0, col = "red")
```



Conclusões do Gráfico Resíduos vs. Leverage

Este gráfico analisa a relação entre os resíduos (diferença entre valores observados e ajustados) e os valores de leverage, que indicam a influência de cada observação no modelo.

---

1. Padrão observado

- **Distribuição geral dos pontos:**
  A maioria dos pontos está concentrada em valores baixos de leverage, indicando que a maioria das observações tem pouca influência no modelo.
- **Alguns valores altos de leverage:**
  Há algumas observações com leverage relativamente alto (próximo de 0.8). Esses pontos representam observações que têm maior influência no ajuste do modelo e podem ser consideradas "observações influentes".
- **Resíduos relativamente distribuídos:**
  Não há um padrão claro de relação entre os resíduos e os valores de leverage, o que é um bom sinal, pois indica que o ajuste não está enviesado por observações com leverage elevado.

---

2. Avaliação de pontos influentes

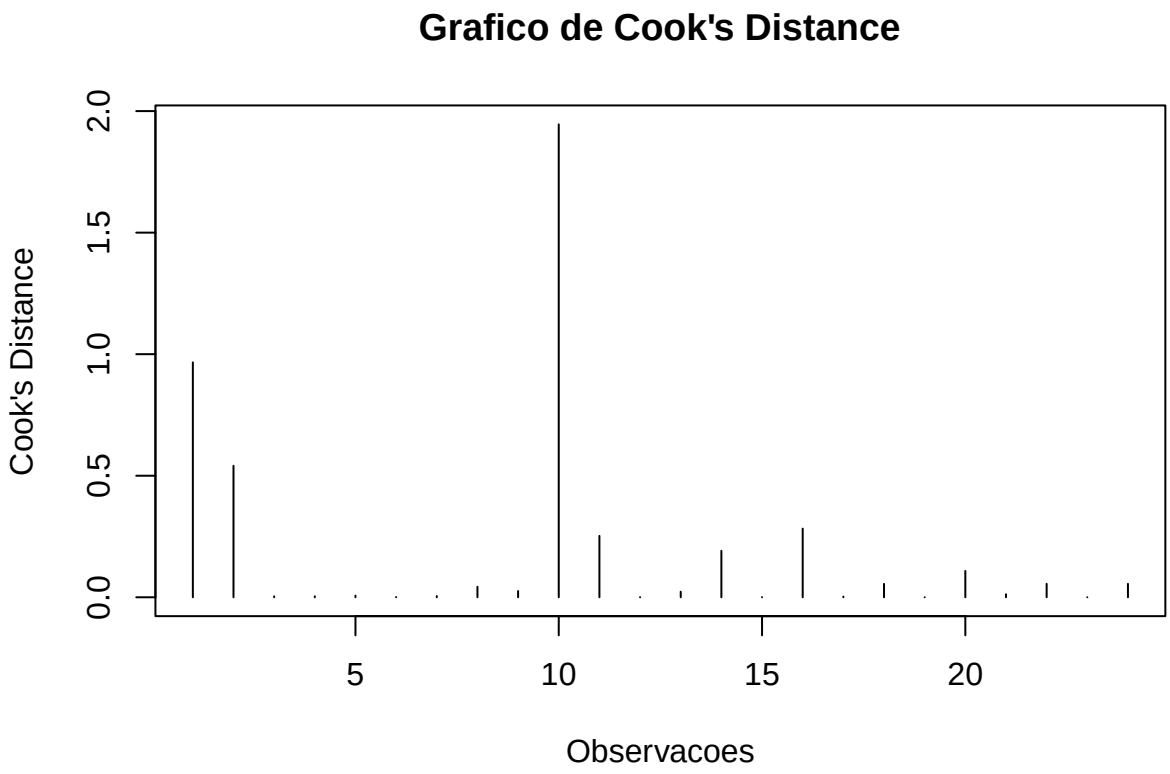
- **Valores de leverage próximos de 1:**
  Nenhum ponto alcança valores de leverage extremamente altos (próximos de 1), mas aqueles com leverage acima de 0.5 devem ser monitorados, pois podem ter um impacto significativo no modelo.
- **Resíduos moderados para altos leverage:**
  Mesmo os pontos com valores mais altos de leverage têm resíduos próximos a zero, indicando que, embora influentes, essas observações não distorcem significativamente o modelo.

Conclusão

O gráfico não apresenta um padrão preocupante entre resíduos e leverage, mas os pontos com valores de leverage mais altos merecem uma análise adicional. A robustez do modelo parece adequada até o momento.

Gráfico de Cook's distance

```
plot(cooks.distance(modelo_cv$finalModel),
     type = "h",
     ylab = "Cook's Distance",
     xlab = "Observacoes",
     main = "Grafico de Cook's Distance")
```




```
which.max(cooks.distance(modelo_cv$finalModel))
```

```
## X10
##  10
```

```
max(cooks.distance(modelo_cv$finalModel))
```

```
## [1] 1.945493
```

Conclusões do Gráfico de Cook's Distance

1. Observação mais influente

- A **observação 10** tem o maior valor de Cook's Distance, com um valor de aproximadamente **1.945**.
- Um valor de Cook's Distance maior que **1** geralmente indica uma observação potencialmente influente, ou seja, que exerce impacto significativo no ajuste do modelo.

2. Impacto da observação 10

- A influência da observação 10 sugere que ela pode estar distorcendo os resultados do modelo. Isso pode ser causado por um outlier ou por leverage alto combinado com resíduos elevados.

3. Outras observações

- A maioria das demais observações apresenta valores de Cook's Distance bem menores do que 1, indicando que sua influência no modelo é mínima.

Conclusão

A observação 10 é altamente influente no modelo, conforme indicado pelo Cook's Distance. É essencial investigar sua origem e testar a robustez do modelo sem ela, para garantir que os resultados não estejam sendo distorcidos de forma significativa.

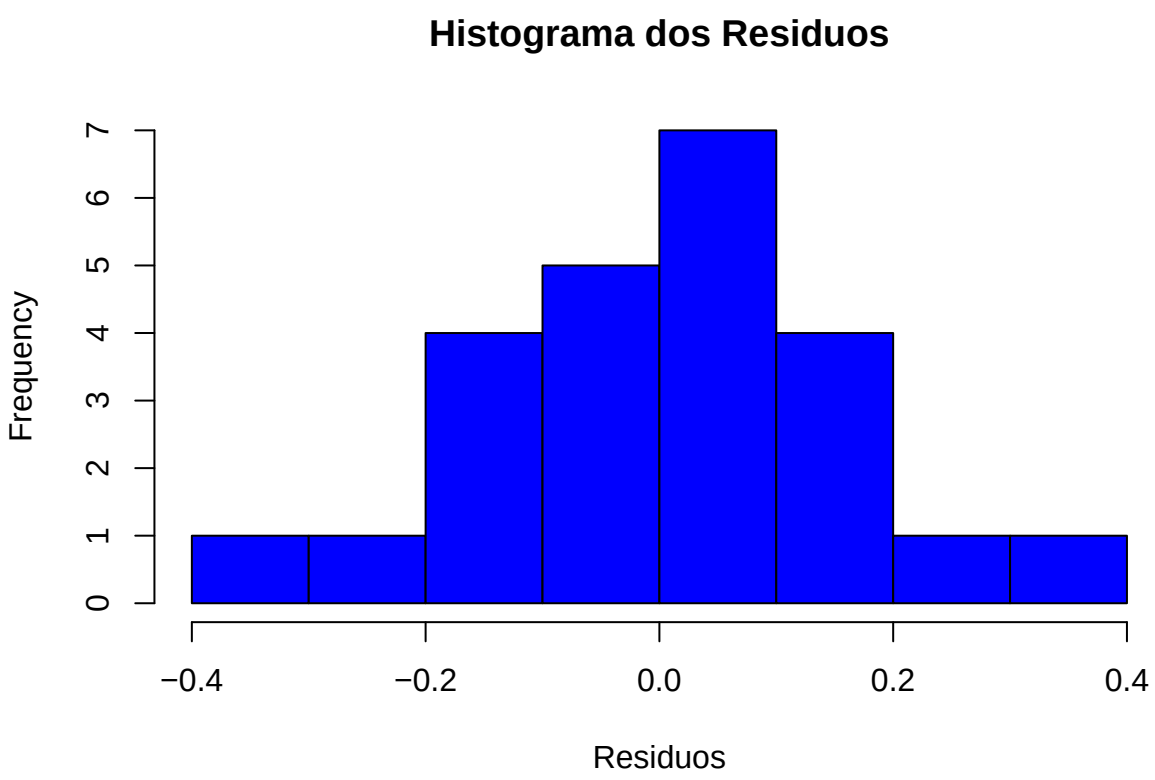
histograma dos residuos

Primeiro, extrair os resíduos do modelo final

```
residuos <- residuals(modelo_cv$finalModel)
```

Criar o histograma dos resíduos

```
hist(residuos, main = "Histograma dos Residuos", xlab = "Residuos", col = "blue")
```



Conclusões do Histograma dos Resíduos

1. **Distribuição Aproximadamente Simétrica:**
   - O histograma dos resíduos apresenta uma distribuição que é aproximadamente simétrica em torno de zero, indicando que os resíduos não possuem viés significativo.
   - Isso é consistente com o pressuposto de normalidade dos resíduos em um modelo de regressão linear.
2. **Resíduos Pequenos e Concentrados:**
   - A maioria dos resíduos está próxima de zero, o que sugere que o modelo ajusta bem os dados.
   - Não há evidências visuais de grandes outliers ou resíduos extremos que possam influenciar negativamente o modelo.
3. **Amplitude dos Resíduos:**
   - Os resíduos estão concentrados em um intervalo de aproximadamente -0,4 a 0,4, o que reflete um erro relativamente pequeno nos valores ajustados pelo modelo.

Recomendações

1. **Confirmar a Normalidade:**
   - Embora o histograma sugira uma distribuição normal, testes formais como o teste de Shapiro-Wilk podem ser usados para validar essa hipótese.

2. **Investigar Outliers:**
    - Apesar de o histograma não mostrar evidências claras de outliers, outras análises (como o Cook's Distance e o gráfico de resíduos vs leverage) já identificaram observações influentes. Isso deve ser explorado em conjunto.
3. **Comparar com Outros Diagnósticos:**
    - Este histograma deve ser analisado junto com o gráfico Q-Q para confirmar a normalidade dos resíduos.

Conclusão

O histograma indica que os resíduos do modelo estão bem distribuídos e próximos de zero, sugerindo que o modelo é adequado. No entanto, deve-se complementar a análise com outras ferramentas diagnósticas para confirmar a qualidade do ajuste e tratar observações influentes.

normalidade

Realizar o teste de Shapiro-Wilk nos resíduos

```
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.99002, p-value = 0.9963
```

Conclusões do Teste de Shapiro-Wilk

1. **Hipótese Nula do Teste de Shapiro-Wilk:**
    - A hipótese nula do teste é que os dados seguem uma distribuição normal.
2. **Resultados do Teste:**
    - **Estatística W = 0.99002:** Próxima de 1, indicando que os resíduos estão bem alinhados com uma distribuição normal.
    - **p-valor = 0.9963:** Muito maior do que o nível de significância típico (como 0,05), indicando que não há evidências suficientes para rejeitar a hipótese nula.
3. **Interpretação:**
    - Os resíduos do modelo podem ser considerados normalmente distribuídos.
    - Isso é consistente com os gráficos de diagnóstico, como o histograma dos resíduos e o gráfico Q-Q, que também sugeriram normalidade.

Conclusão

O teste de Shapiro-Wilk confirma que os resíduos seguem uma distribuição normal. Isso reforça a adequação do modelo de regressão linear, pois a normalidade dos resíduos é um pressuposto essencial para validar as inferências do modelo. Não há necessidade de transformações adicionais ou ajustes para tratar a normalidade.

homocedasticidade

Aplicar o teste de Breusch-Pagan para verificar a homocedasticidade dos resíduos

```
ncvTest(modelo_cv$finalModel)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.98516, Df = 1, p = 0.025566
```

Conclusões do Teste de Variância Constante (ncvTest)

1. **Objetivo do Teste:**
   - O teste de variância constante avalia a homocedasticidade, ou seja, se os resíduos possuem variância constante em relação aos valores ajustados do modelo.
   - A **hipótese nula** do teste é que a variância dos resíduos é constante (homocedasticidade).
2. **Resultados do Teste:**
   - **Chisquare = 4.98516** com **p-valor = 0.025566**.
   - O p-valor é menor que o nível de significância comum (0.05), indicando que há evidências estatísticas para rejeitar a hipótese nula.
3. **Interpretação:**
   - A rejeição da hipótese nula sugere que os resíduos não possuem variância constante (heterocedasticidade).
   - Isso viola um dos pressupostos fundamentais do modelo de regressão linear e pode indicar que os erros do modelo não estão distribuídos de forma uniforme ao longo do intervalo dos valores ajustados.

autocorrelacao

Realizar o teste de Durbin-Watson nos resíduos

```
durbinWatsonTest(residuos)
```

```
## [1] 1.079303
```

Conclusões sobre o Teste de Durbin-Watson

1. **Objetivo do Teste:**
   - O teste de Durbin-Watson avalia a presença de autocorrelação nos resíduos de um modelo de regressão linear.
   - A **hipótese nula (H)** do teste é que **não existe autocorrelação nos resíduos** (os resíduos são independentes).
2. **Resultados do Teste:**
   - O valor estatístico obtido foi **1.079303**.
3. **Interpretação:**
   - O valor da estatística de Durbin-Watson varia entre 0 e 4:
     - Um valor próximo de **2** indica **ausência de autocorrelação**.
     - Valores menores que **2** indicam **autocorrelação positiva**.
     - Valores maiores que **2** indicam **autocorrelação negativa**.
   - O valor obtido (**1.079**) está bem abaixo de **2**, sugerindo **autocorrelação positiva** nos resíduos.
4. **Implicações:**
   - A presença de autocorrelação positiva indica que os resíduos não são completamente independentes.
   - Esse comportamento viola outro pressuposto da regressão linear, comprometendo a validade das inferências estatísticas (como os intervalos de confiança e os testes de hipóteses).

multicolinearidade

```
vif(modelo_cv$finalModel)
```

```
##    ln_area sanitario2 sanitario3 sanitario4        ano  ln_fiscal
##   4.570712   3.395598   3.293352   5.594772   1.375207   1.241184
```

Análise de Colinearidade com o VIF

1. **Objetivo do VIF:**
   - O **VIF (Variance Inflation Factor)** mede o grau de multicolinearidade entre as variáveis explicativas do modelo.
   - Valores elevados de VIF indicam que uma variável está altamente correlacionada com outras, o que pode prejudicar a interpretação dos coeficientes.
2. **Resultados Obtidos:**
   - **ln_area**: 4.57
   - **sanitario2**: 3.40
   - **sanitario3**: 3.29
   - **sanitario4**: 5.59

- **ano**: 1.38
- **ln_fiscal**: 1.24

3. **Critérios de Interpretação:**
   - Geralmente:
     - **VIF < 5**: Colinearidade baixa, aceitável.
     - **VIF entre 5 e 10**: Colinearidade moderada, pode exigir atenção.
     - **VIF > 10**: Colinearidade alta, problemático, deve ser corrigido.
4. **Conclusões:**
   - **ln_area** e **sanitario4** apresentam os maiores VIFs:
     - **ln_area (4.57)**: Está próximo de 5, indicando colinearidade moderada, mas ainda aceitável.
     - **sanitario4 (5.59)**: Exibe colinearidade moderada, levemente acima do limite aceitável.
   - As demais variáveis possuem **VIFs baixos (< 5)**, indicando colinearidade baixa ou inexistente.
5. **Implicações:**
   - A presença de colinearidade moderada para **sanitario4** pode dificultar a interpretação do coeficiente desta variável.
   - Apesar disso, a multicolinearidade geral do modelo não parece ser um problema grave.

GVLMA

```
gvlma(modelo_cv$finalModel)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)      ln_area   sanitario2   sanitario3   sanitario4          ano
##   -28.12688      0.90668      0.20038      0.52642      0.43250      0.01782
##   ln_fiscal
##     0.34071
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = modelo_cv$finalModel)
##
##                      Value  p-value                   Decision
## Global Stat        9.077360 0.059195    Assumptions acceptable.
## Skewness           0.006675 0.934886    Assumptions acceptable.
## Kurtosis           0.028913 0.864979    Assumptions acceptable.
## Link Function      6.944200 0.008409 Assumptions NOT satisfied!
## Heteroscedasticity 2.097572 0.147533    Assumptions acceptable.
```

Análise dos Resultados do Teste GVLMA

O pacote `gvlma` fornece uma avaliação abrangente das suposições de um modelo linear, com base em quatro aspectos principais:

1. **Global Stat:** Verifica todas as suposições do modelo como um todo.
2. **Skewness:** Testa a simetria dos resíduos.
3. **Kurtosis:** Testa a adequação da distribuição normal dos resíduos (grau de achatamento).
4. **Link Function:** Testa se a especificação do modelo linear é correta.
5. **Heteroscedasticity:** Testa a homogeneidade da variância dos resíduos.

Resultados Obtidos:

1. **Global Stat (Estatística Global):**
   - **Valor:** 9.077
   - **p-valor:** 0.059
   - **Decisão:** As suposições são consideradas aceitáveis no nível de significância de 5%. Embora o valor do teste seja próximo ao limite, o modelo não viola grosseiramente as suposições.
2. **Skewness (Assimetria):**
   - **Valor:** 0.007
   - **p-valor:** 0.935
   - **Decisão:** Os resíduos não apresentam assimetria significativa. Suposição aceita.
3. **Kurtosis (Curtose):**
   - **Valor:** 0.029
   - **p-valor:** 0.865
   - **Decisão:** A distribuição dos resíduos tem uma curtose adequada. Suposição aceita.
4. **Link Function (Especificação do Modelo):**
   - **Valor:** 6.944
   - **p-valor:** 0.008
   - **Decisão: A suposição de especificação correta do modelo não é satisfeita.**
     - Isso sugere que o modelo linear pode não estar capturando adequadamente a relação entre as variáveis. Pode haver termos não lineares ou interações que foram omitidos.
5. **Heteroscedasticity (Heterocedasticidade):**
   - **Valor:** 2.098
   - **p-valor:** 0.148
   - **Decisão:** Não há evidências de heterocedasticidade. A suposição de variância constante dos resíduos é aceita.

Conclusões:

1. **Assunções em Geral:**
   - A maioria das suposições do modelo linear é aceita, indicando que o modelo tem uma boa aderência às premissas básicas de linearidade, normalidade dos resíduos e homocedasticidade.

2. **Problema na Especificação do Modelo:**
   - O teste de **Link Function** indica que o modelo linear pode estar incompleto ou mal especificado. Isso sugere que:
     - Pode haver relações não lineares entre as variáveis.
     - Algumas variáveis ou interações importantes podem estar ausentes.

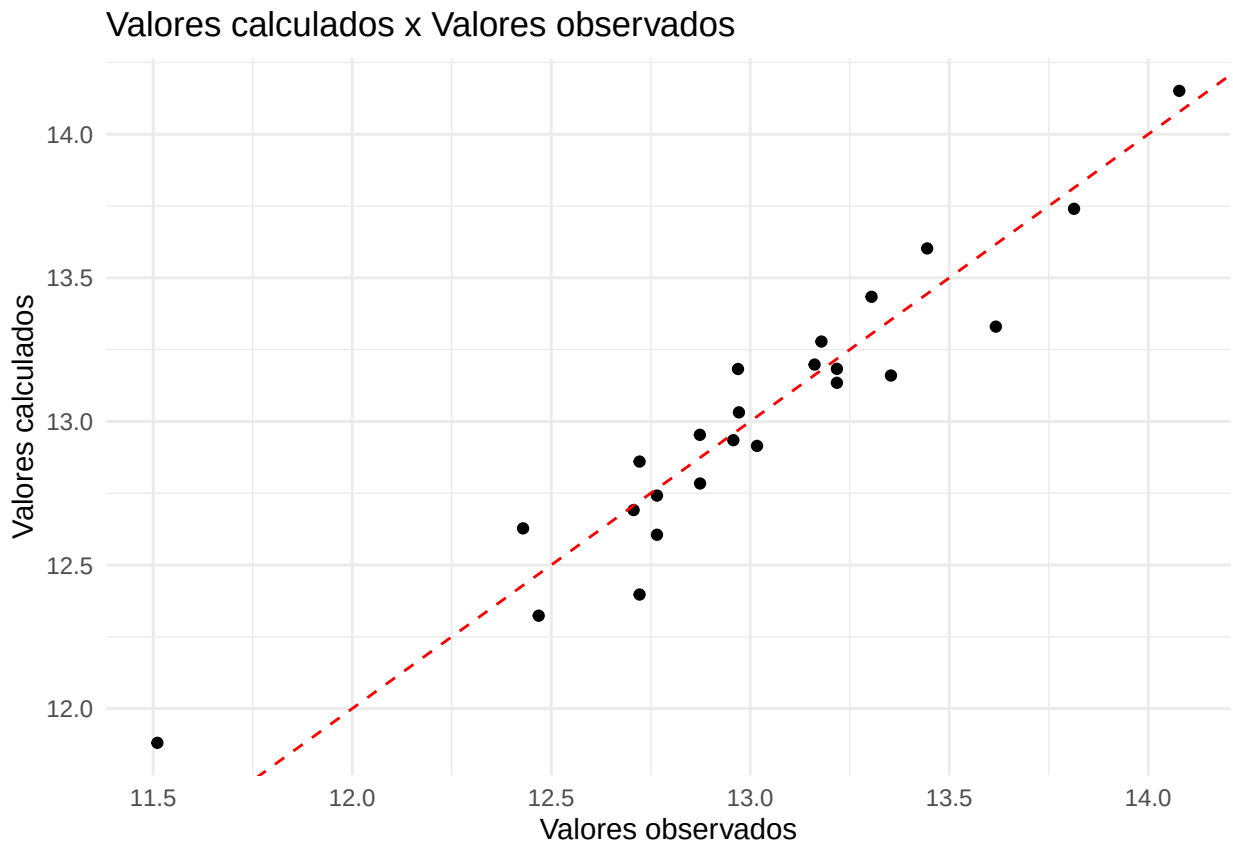
Calculado x Observado

Calculando as previsões e adicionando ao dataframe original

```
apartamento$predicted <- predict(modelo_cv, newdata = apartamento)
```

Criando o gráfico

```
ggplot(data = apartamento) +
  geom_point(aes(x = ln_valor, y = predicted)) +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = 'Valores observados', y = 'Valores calculados', title = 'Valores calculados x Valores observa
  theme_minimal()
```



Análise do Gráfico: Valores Calculados x Valores Observados

O gráfico apresenta os valores observados (`ln_valor`) no eixo X e os valores previstos pelo modelo (`predicted`) no eixo Y. A linha vermelha tracejada representa a diagonal $y = x$, que é a linha de perfeição — ou seja, os pontos que caem exatamente nessa linha indicariam previsões perfeitas pelo modelo.

Conclusões do Gráfico

1. **Aderência ao Modelo:**
   - A maior parte dos pontos está próxima da linha de perfeição, indicando que o modelo tem um bom desempenho na previsão dos valores log-transformados de *ln_valor*.
   - Isso é consistente com a alta R^2 ajustada (~0.88) observada na análise do modelo, sugerindo que a maior parte da variação nos valores observados é explicada pelas variáveis preditoras no modelo.
2. **Distribuição dos Erros:**
   - Não há grandes desvios sistemáticos visíveis. Os pontos não parecem se desviar consistentemente para cima ou para baixo da linha, o que indica que o modelo captura bem a relação entre as variáveis.
   - Pequenas discrepâncias nos extremos sugerem a presença de alguns valores menos bem ajustados, possivelmente outliers ou dados com maior incerteza.
3. **Verificação de Tendências Não Lineares:**
   - Não há um padrão curvo ou sistemático nos pontos, o que reforça que o modelo linear é uma boa escolha para esses dados transformados.

Próximos Passos ou Considerações

1. **Avaliação de Outliers:**
   - Embora o gráfico mostre um ajuste geral bom, é importante avaliar mais detalhadamente os pontos que estão mais afastados da linha de perfeição. Esses pontos podem indicar outliers ou casos onde o modelo não está performando bem.
2. **Validação Cruzada:**
   - Já foi realizado um processo de validação cruzada, que indicou um RMSE de 0.27. Isso também é refletido no gráfico, que confirma que as previsões estão alinhadas com os valores reais.
3. **Possível Melhoria no Modelo:**
   - Para reduzir ainda mais os desvios, pode-se explorar a inclusão de novas variáveis, transformações não lineares, ou interações entre variáveis.

Resumo

O gráfico reforça que o modelo ajustado é robusto e explica bem os dados observados. Apesar disso, algumas discrepâncias menores indicam que melhorias ainda podem ser exploradas para refinar a precisão do modelo.