Objetivo do projeto: desenvolver um modelo de regressão linear para prever o preço das casas com base nas variáveis fornecidas:

- SalePrice: Preço de venda em dólares
- Basement_Area: Área do porão em pés quadrados
- Lot_Area: Tamanho do lote em pés quadrados
- Heating_QC: Qualidade e condição do aquecimento
- Season_Sold: Estação quando a casa foi vendida
- Gr_Liv_Area: Área acima do nível do solo em pés quadrados
- Garage_Area: Tamanho da garagem em pés quadrados
- Deck_Porch_Area: Área total de decks e varandas em pés quadrados
- Age_Sold: Idade da casa quando vendida, em anos
- Bedroom_AbvGr: Quartos acima do nível do solo
- Total_Bathroom: Número total de banheiros (meio banheiro contado como 10%)

Heating_QC e Season_Sold são variáveis categóricas, então temos que converter em dummies.
Heating_QC: variável nominal ordinal (Ex > Gd > TA > Fa > Po)
Representa a qualidade e condição do sistema de aquecimento em uma escala ordinal.

- **Ex**: **Excellent** (Excelente) – em ótimo estado.
- **Gd**: **Good** (Bom) – em bom estado.
- **TA**: **Typical/Average** (Típico ou Médio) – na média.
- **Fa**: **Fair** (Regular) – abaixo da média.
- **Po**: **Poor** (Ruim) – em más condições.

Season_Sold: variável categórica que indica a estação em que foi vendida. Apesar de ser numérica representa uma informação associada a estação do ano.

Variáveis com valores muito elevados dificultando a análise e por esse motivo foi aplicada uma transformação (logaritmo dos valores)

Outra decisão foi com relação as diversas áreas existentes no imóvel que possuem uma distribuição assimétrica e com muitos outliers, com isso, decidi trabalhar com uma nova variável (área total) que é a soma das áreas (area_total = Basement_Area + Lot_Area + Gr_Liv_Area + Garage_Area + Deck_Porch_Area) e não utilizá-las do modelo para evitar multicolinearidade.
Dessa forma obtive a matrix de correlação sem indício de multicolinearidade.
Análise univariada e verificação de inconsistências:

```
> summary(precos)
   SalePrice      Basement_Area     Lot_Area       Heating_QC         Season_Sold     Gr_Liv_Area    Garage_Area
 Min.   : 12789   Min.   :   0   Min.   :  1300   Length:2928        Min.   :1.000   Min.   : 334   Min.   :   0.0
 1st Qu.:129500   1st Qu.: 793   1st Qu.:  7441   Class :character   1st Qu.:2.000   1st Qu.:1126   1st Qu.: 320.0
 Median :160000   Median : 990   Median :  9444   Mode  :character   Median :3.000   Median :1442   Median : 480.0
 Mean   :180841   Mean   :1052   Mean   : 10150                      Mean   :2.608   Mean   :1500   Mean   : 472.9
 3rd Qu.:213500   3rd Qu.:1302   3rd Qu.: 11556                      3rd Qu.:3.000   3rd Qu.:1742   3rd Qu.: 576.0
```

```
Max.   :755000   Max.   :6110   Max.   :215245                    Max.   :4.000   Max.   :5642   Max.   :1488.0
Deck_Porch_Area     Age_Sold      Bedroom_AbvGr   Total_Bathroom
Min.   :   0.0   Min.   : -1.00   Min.   :0.000   Min.   :0.400
1st Qu.:  22.0   1st Qu.:  7.00   1st Qu.:2.000   1st Qu.:1.100
Median : 140.0   Median : 34.00   Median :3.000   Median :2.000
Mean   : 159.9   Mean   : 36.41   Mean   :2.855   Mean   :2.042
3rd Qu.: 247.0   3rd Qu.: 54.00   3rd Qu.:3.000   3rd Qu.:2.100
Max.   :1424.0   Max.   :136.00   Max.   :8.000   Max.   :6.200
```

SalePrice (Preço de Venda)
Descrição: preço de venda das casas.
Distribuição: A mediana é menor que a média, indicando uma assimetria à direita (presença de outliers ou valores extremos altos).

Basement_Area (Área do Porão)
Descrição: Área total do porão em feet quadrados.
Distribuição: A mediana é próxima da média, sugerindo uma distribuição relativamente simétrica.

Lot_Area (Área do Terreno)
Descrição: Área total do terreno em feet quadrados.
Distribuição: A média é maior que a mediana, indicando uma assimetria à direita (possíveis outliers ou terrenos muito grandes).

Heating_QC (Qualidade do Aquecimento)
Descrição: Qualidade do sistema de aquecimento (variável categórica).

Season_Sold (Estação da Venda)
Descrição: Estação do ano em que a casa foi vendida
Mais provável? 1: Inverno, 2: Primavera, 3: Verão, 4: Outono
Distribuição: A maioria das vendas ocorre mediana = 3

Gr_Liv_Area (Área Habitável)
Descrição: Área habitável da casa em feet quadrados.
Distribuição: A média e a mediana são próximas, sugerindo uma distribuição simétrica.

Garage_Area (Área da Garagem)
Descrição: Área da garagem em metros quadrados.
Distribuição: A mediana é próxima da média, indicando uma distribuição simétrica.

Deck_Porch_Area (Área do Deck/Varanda)
Descrição: Área do deck ou varanda em feet quadrados.
Distribuição: A média é maior que a mediana, indicando uma assimetria à direita (possíveis outliers).

Age_Sold (Idade da Casa na Venda)
Descrição: Idade da casa no momento da venda (em anos).
Mínimo: -1 ano (possível erro ou casa ainda em construção).
Distribuição: A mediana é próxima da média, sugerindo uma distribuição simétrica.

Bedroom_AbvGr (Número de Quartos)
Descrição: Número de quartos acima do nível do solo.
Análise:
Mínimo: **0 quartos (possível erro ou casa sem quartos).**
Distribuição: A média e a mediana são próximas, indicando uma distribuição simétrica.

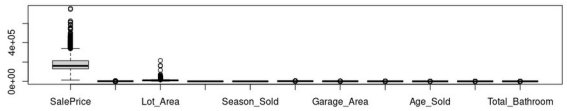
Total_Bathroom (Número Total de Banheiros)
Descrição: Número total de banheiros
Análise:
Mínimo: **0,4 banheiros (possível erro ou casa sem banheiros).**
istribuição: A média e a mediana são próximas, sugerindo uma distribuição simétrica.

Dicionário de Dados

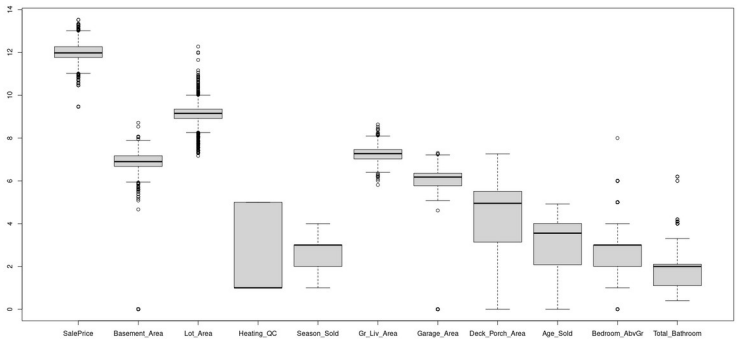
Nome da Variável	Descrição	Tipo de Dado	Unidade/Métrica		Observações
SalePrice	Preço de venda da casa	Numérico	US$		Variável dependente (target).
Basement_Area	Área total do porão	Numérico	Feets quadrados (f²)		Valor mínimo = 0 (sem porão).
Lot_Area	Área total do terreno	Numérico	Feets quadrados (f²)		Possíveis outliers (valores altos).
Heating_QC	Qualidade do sist de aquecimento	Categórico	Texto		Precisa ser convertido em fator.
Season_Sold	Estação do ano	Categórico			
Gr_Liv_Area	Área habitável da casa	Numérico	Feets quadrados (f²)		Distribuição simétrica.
Garage_Area	Área da garagem	Numérico	Feets quadrados (f²)		Valor mínimo = 0 (sem garagem).
Deck_Porch_Area	Área do deck ou varanda	Numérico	Feets quadrados (f²)		Valor mínimo = 0 (sem deck/varanda).

Age_Sold	Idade casa (momento da venda)	Numérico	Anos	Valor mínimo = -1 (possível erro).
Bedroom_AbvGr	Número de quartos	Numérico	Quantidade	Valor mínimo = 0 (possível erro).
Total_Bathroom	Número total de banheiros	Numérico	Quantidade	Inclui banheiros parciais (0,5).

Boxplot

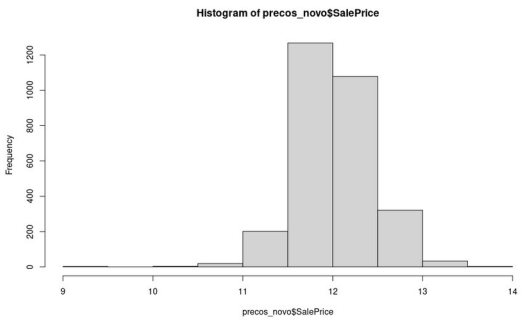


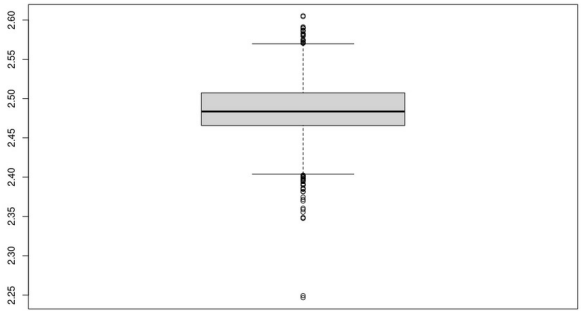
Necessidade de aplicar transformação (logaritmo) para poder analisar.
Presença de outliers em diversas variáveis.



A distribuição da variável SalesPrice – **log(SalesPrice)**: distribuição simétrica

```
> summary(log(precos_novo$SalePrice))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.247   2.466   2.483   2.486   2.507   2.605
```

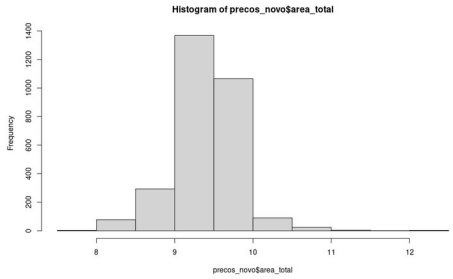


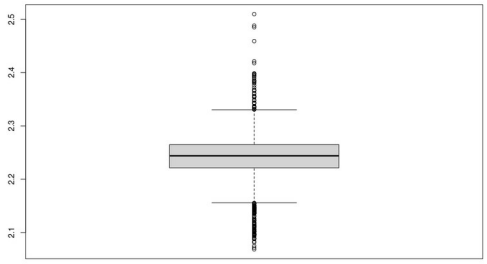


Decisões tomadas:
Criar e trabalhar com uma nova variável: área total do imóvel ao invés das áreas separadamente.
area_total = Basement_Area + Lot_Area + Gr_Liv_Area + Garage_Area + Deck_Porch_Area
area_total: a média e a mediana são próximas, indicando uma distribuição simétrica. Presença de outliers superiores e inferiores.
Min. 1st Qu. Median Mean 3srd Qu. Max.

7.912 9.219 9.430 9.407 9.630 12.301

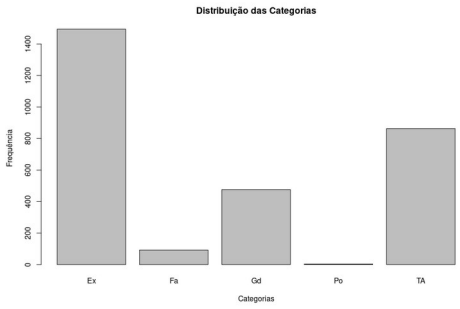
A distribuição da variável area_total – **log(area_total)**: distribuição simétrica. Presença de outliers superiores e inferiores.

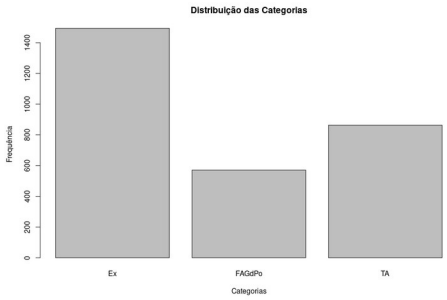




Avaliando o volume das variáveis categóricas

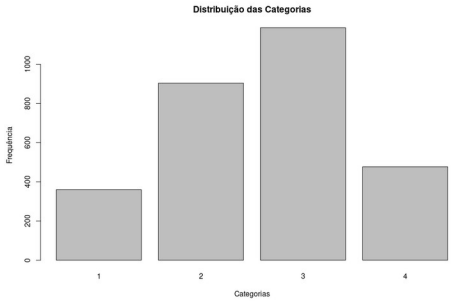
Heating_QC

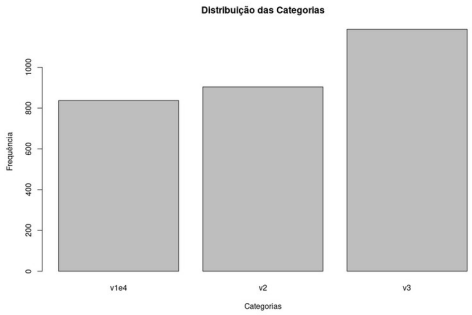




Ajustando o volume, pois está muito discrepante.
Categorias: EX, (FA+Gd+Po) e TA
EX: casela de referência por ser o maior volume

Season_Sold

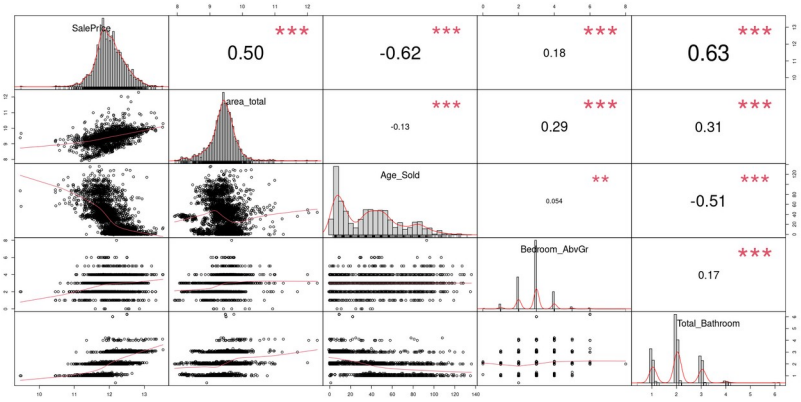




Ajustando o volume
V3: casela de referência

Análise bivariada

Matriz de correlações



Não existe um indicativo de multicolinearidade entre as variáveis escolhidas para integrar o modelo: SalePrice,area_total,Age_sold,Bedroom_AbvGr, Total_Bathroom, Heating_QC e Season_Sold (as duas últimas são categóricas e com isso, serão criadas variáveis dummies)

Criação do modelo nulo e o modelo completo para usar o step(forward)

```
#modelo nulo
lm_precos_nulo <- lm(SalePrice ~ 1, data=precos_dummies)

#modelo completo
lm_precos_full <- lm(SalePrice ~ area_total + Age_Sold + Bedroom_AbvGr + Total_Bathroom + heating_agrupadaFAGdPo + heating_agrupadaTA + season_agrupadav2 + season_agrupadav1e4, data=precos_dummies)

# Step partindo do modelo nulo até o modelo completo
forw <- step(lm_precos_nulo, scope=list(lower=lm_precos_nulo, upper=lm_precos_full), direction = "forward")
summary(forw)
```

Melhor resultado: AIC=-8392.27

```
SalePrice ~ Total_Bathroom + Age_Sold + area_total + heating_agrupadaTA + heating_agrupadaFAGdPo + Bedroom_AbvGr
```

```
lm(formula = SalePrice ~ Total_Bathroom + Age_Sold + area_total + 
heating_agrupadaTA + heating_agrupadaFAGdPo + Bedroom_AbvGr, data = 
precos_dummies)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.3039 -0.1363 -0.0037  0.1319  1.0142 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              8.843635   0.105601  83.746  < 2e-16 ***
Total_Bathroom           0.152576   0.007168  21.285  < 2e-16 ***
Age_Sold                -0.004673   0.000183 -25.537  < 2e-16 ***
```

```
area_total              0.321919   0.011762  27.369  < 2e-16 ***
heating_agrupadaTA     -0.178120   0.011248 -15.835  < 2e-16 ***
heating_agrupadaFAGdPo -0.121491   0.012613  -9.632  < 2e-16 ***
Bedroom_AbvGr           0.029452   0.005656   5.207 2.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2383 on 2921 degrees of freedom
Multiple R-squared:  0.6588,    Adjusted R-squared:  0.6581
F-statistic: 939.8 on 6 and 2921 DF,  p-value: < 2.2e-16
```

Observações:

Todos os coeficientes são estatisticamente significativos (p-value < 0.001), extremamente significativo para todas as variáveis e nenhuma variável pode ser considerada irrelevante.

Coeficientes
Impacto no preço (log) por unidade de variação:
Impacto positivo:
area_total: +0.322 (maior impacto positivo)
Total_Bathroom: +0.153
Bedroom_AbvGr: +0.029

Impacto negativo:
Age_Sold: -0.005
heating_agrupadaTA: -0.178
heating_agrupadaFAGdPo: -0.121

Qualidade do modelo:
R-squared: 0.6588 (65.88% da variância explicada)
Adjusted R-squared: 0.6581 (muito próximo do R-squared)

Resíduos
Simetria próxima de zero (Median: -0.0037)
Amplitude: -2.3039 a 1.0142
Erro padrão residual: 0.2383 (baixo)

O modelo parece bem ajustado, pois explica 66% da variabilidade do preço de venda com variáveis estatisticamente significativas.
Podem haver variáveis importantes ausentes, já que 34% da variação ainda não é explicada.

Calculando VIF

ols_vif_tol(lm_bodyfat_final_AIC)

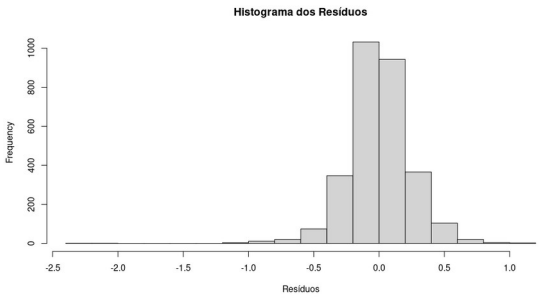
```
> ols_vif_tol(lm_precos_aic)
                  Variables Tolerance      VIF
1         Total_Bathroom 0.6591966 1.516998
2               Age_Sold 0.6315267 1.583464
3             area_total 0.8413160 1.188614
4     heating_agrupadaTA 0.7372998 1.356300
5 heating_agrupadaFAGdPo 0.7764692 1.287881
6          Bedroom_AbvGr 0.8847687 1.130239
```

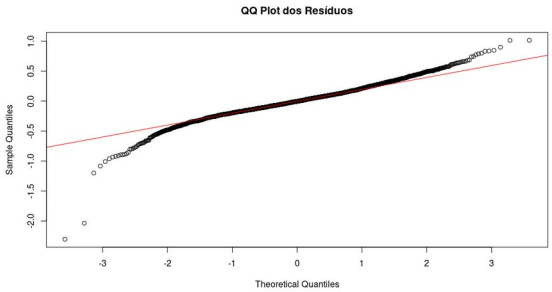
Nenhum valor de VIF acima de 2, um forte indício de ausência de multicolinearidade

Os resíduos não seguem uma Distribuição Normal: (95% de nível de confiança)

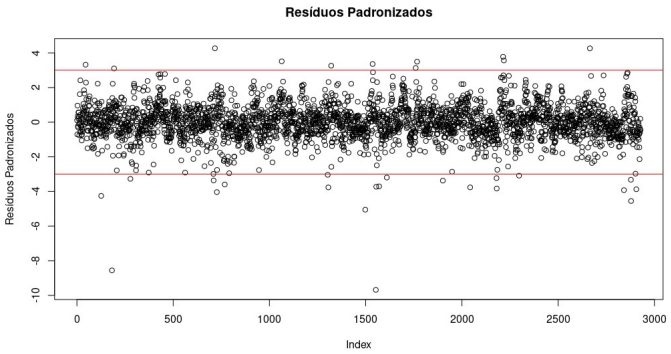
```
> # Teste de normalidade (Shapiro-Wilk)
> resultado <- shapiro.test(residuals(lm_precos_aic))
> print(resultado$p.value)
[1] 3.700888e-28
```

Histograma e normalidade (QQ)





Resíduos padronizados distribuídos não estão aleatoriamente em torno de 0.



Teste de homogeneidade de Variância dos resíduos indica heterocedasticidade

```
> bptest(lm_precos_aic)
```

```
  studentized Breusch-Pagan test

data:  lm_precos_aic
BP = 99.299, df = 6, p-value < 2.2e-16
Hipótese alternativa (H1): Os resíduos possuem variância não constante
(heterocedasticidade).P-valor < 0.05
```

Qualidade do modelo

MSE (Erro Quadrático Médio) = 0.0566
RMSE (Raiz do Erro Quadrático Médio) = 0.2380

Como SalePrice está transformado em logaritmo (log(SalePrice), o RMSE precisa ser interpretado no domínio original dos preços.

O RMSE 0.2380 representa o erro médio dos logaritmos dos preços.

Interpretação "desfazendo" a transformação

O erro percentual médio: erro_perc <- exp(rmse) - 1

Erro percentual médio aproximadamente 26,9%
Isso significa que, em média, as previsões do modelo diferem do valor real de SalePrice em aproximadamente 26,9%.

Conclusões:

Modelo é razoavel: possui um bom ajuste inicial, com um R-squared de cerca de 65%, erros médios razoáveis e sem problemas significativos de multicolinearidade.
A presença de heterocedasticidade sugere que a variância dos resíduos não é constante, talvez uma outra transformação possa melhorar.
A ausência de normalidade dos resíduos é um problema que pode afetar a robustez dos testes de hipóteses e a confiabilidade do modelo, então seria interessante investigar transformações adicionais ou modelagem não linear.
Erro Percentual Médio: O modelo apresenta um erro percentual médio relativamente alto.