Universidade de Brasília
Programa de Pós-graduação em Computação Aplicada – PPCA
Análise Estatística de Dados e Informações (AEDI)

Relatório - Regressão Linear

Aluno: João Robson Santos Martins

Professor:
João Gabriel de Moraes Souza

15 de dezembro de 2024

1 Introdução

Este relatório apresenta uma análise estatística de um *dataset* de imóveis da cidade de Ames, no Iowa, EUA. O dado foi estruturado por [1] e sua documentação pode ser vista aqui.

A ideia é criar um modelo de regressão capaz de predizer o preço de venda médio desses imóveis com base em cinco características de cada imóvel: ano de construção, área total, bairro, número de vagas e qualidade geral.

O *notebook* com a análise completa pode ser encontrado aqui.

2 Regressão Linear

A regressão linear [6] é uma técnica estatística usada para modelar a relação entre uma variável dependente (também chamada de variável de resposta ou variável explicada) e uma ou mais variáveis independentes (também chamadas de variáveis explicativas ou preditoras). O objetivo é encontrar a melhor linha reta (ou plano, no caso de mais de uma variável independente) que descreve essa relação de forma aproximada.

2.1 Tipos de Regressão Linear

Regressão Linear Simples: Quando há apenas uma variável independente. O modelo é representado por uma equação do tipo:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

Onde:

- Y é a variável dependente.
- X é a variável independente.
- β_0 é o intercepto, ou seja, o valor de Y quando $X = 0$
- β é o coeficiente de regressão, que indica o impacto de X sobre Y
- ϵ é o erro ou resíduo, que representa a diferença entre o valor real de Y e o valor previsto pelo modelo.

Regressão Linear Múltipla: Quando há duas ou mais variáveis independentes. A equação do modelo é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{2}$$

Onde $X_1, X_2, ..., X_n$ são as variáveis independentes e $\beta_0, \beta_1, ..., \beta_n$ são os coeficientes correspondentes.

2.2 Pressupostos da regressão linear

Para que os resultados da regressão linear sejam válidos, alguns pressupostos precisam ser atendidos:

- Linearidade: A relação entre as variáveis independentes e a dependente deve ser linear.
- Homoscedasticidade: A variância dos resíduos deve ser constante para todos os valores das variáveis independentes.
- Normalidade dos resíduos: Os resíduos devem seguir uma distribuição normal.
- Não multicolinearidade: As variáveis independentes não devem ser altamente correlacionadas entre si.

2.3 Variáveis Escolhidas

As variáveis independentes escolhidas para a regressão são:

- Número de vagas na garagem (**Garage Cars**);
- Ano de construção (**Year Built**);
- Bairro do imóvel (**Neighborhood**).
- Área total (**Total SF ("Gr Liv Area"+ "Total Bsmt SF")**);
- Qualidage geral (**Overall Qual**).

As escolhas foram feitas pelos seguintes motivos:

1. Variáveis Numéricas com Correlação mais Alta em Relação ao Preço de Venda

Existem 12 variáveis com correlação positiva superior a 0.5 em relação ao preço de venda, listadas abaixo:

- *Overall Qual* (0.799262)
- *Total SF* (0.790073)
- *Gr Liv Area* (0.706780)
- *Garage Cars* (0.647877)
- *Garage Area* (0.640401)
- *Total Bsmt SF* (0.632280)
- *1st Flr SF* (0.621676)
- *Year Built* (0.558426)
- *Full Bath* (0.545604)
- *Year Remod/Add* (0.532974)
- *Garage Yr Blt* (0.526965)
- *Mas Vnr Area* (0.508285)

Entre essas, 8 variáveis possuem correlação bastante alta entre si, como ilustrado abaixo:

- *Garage Cars* e *Garage Area* (0.9)
- *Garage Yr Blt* e *Year Built* (0.83)
- *Gr Liv Area* e *TotRms AbvGrd* (0.8)
- *1st Flr SF* e *Total Bsmt SF* (0.8)

Além disso, a variável *Gr Liv Area* é a soma de *1st Flr SF*, *2nd Flr SF* e *Low Qual Fin SF*. Para simplificar o modelo, podemos remover 3 dessas variáveis, ficando com as seguintes:

- *Overall Qual* (0.799262)
- *Total SF* (0.790073)
- *Gr Liv Area* (0.706780)
- *Garage Cars* (0.647877)

- *Total Bsmt SF* (0.632280)
- *Year Built* (0.558426)
- *Full Bath* (0.545604)
- *Year Remod/Add* (0.532974)
- *Mas Vnr Area* (0.508285)

2. Simplificação do Modelo

A área total (*Total SF*), que é a soma de *Gr Liv Area* e *Total Bsmt SF*, possui uma correlação de quase 0.8 com o preço de venda. Dessa forma, podemos remover mais duas variáveis:

- *Overall Qual* (0.799262)
- *Total SF* (0.790073)
- *Garage Cars* (0.647877)
- *Year Built* (0.558426)
- *Full Bath* (0.545604)
- *Year Remod/Add* (0.532974)
- *Mas Vnr Area* (0.508285)

Com isso, restam 3 variáveis com correlação superior a 0.6 em relação ao preço. O *ano de construção*, por exemplo, é uma escolha óbvia, visto que há uma grande variação na idade dos imóveis, como será mostrado na análise abaixo.

3. Influência do Bairro no Preço

A variável *bairro* (*Neighborhood*) também costuma ter uma influência significativa no preço dos imóveis. As análises subsequentes confirmam essa hipótese, evidenciando a importância do bairro na determinação do preço de venda.

3 Análise exploratória

Em suma, a análise exploratória dos dados destacou os seguintes padrões presentes no dado:

- Muitas *features* possuem uma alta contagem de registros com valor 0 ("Pool Area"e "Mas Vnr Area", por exemplo). Isso possivelmente se deve ao fato de imóveis em que esse critério não se aplica. Ou seja, imóveis sem piscina tem área da piscina sendo 0, por exemplo.
- Os preços de venda possuem uma distribuição assimétrica à direita, contendo muitos *outliers* à direita (valores mais caros), conforme a figura 1.

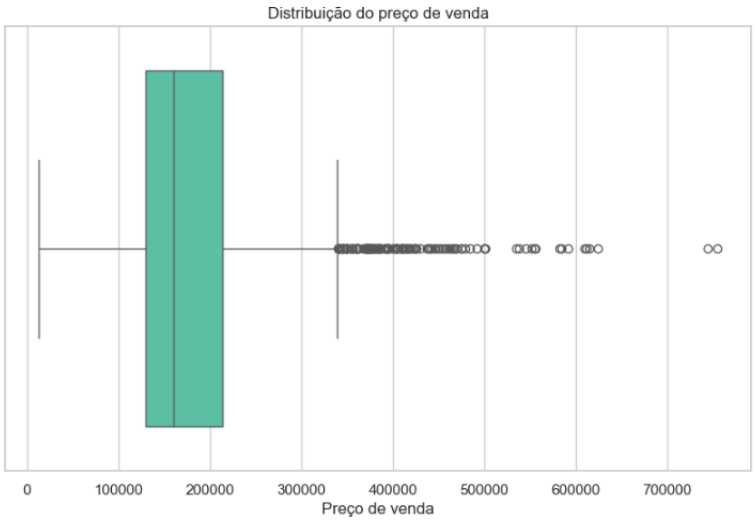

Figura 1: Boxplot dos preços de venda

3.1 Correlação entre *features*

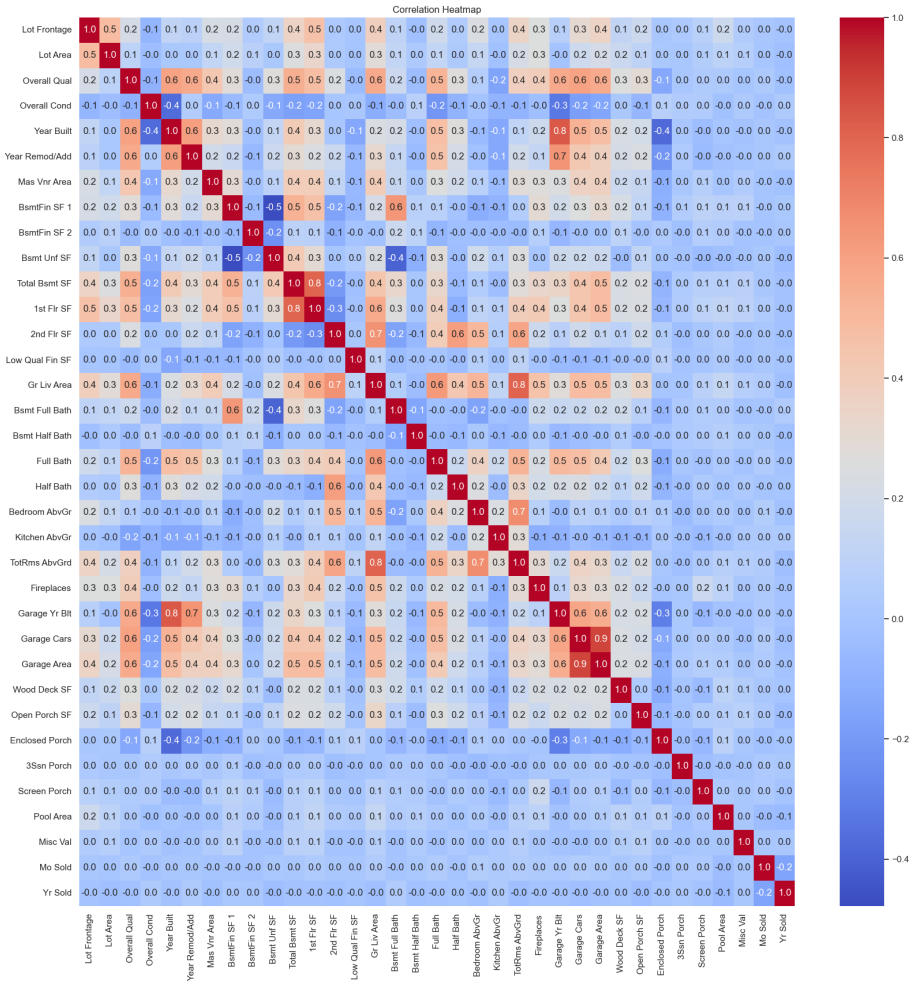

Figura 2: Correlação das *features* com preço de venda

Na análise de correlação entre *features*, conforme a figura 3, os seguintes pontos se destacam:

- A qualidade geral ("Overall Qual") está fortemente correlacionada com o ano de construção e a capacidade da garagem.
- A área da piscina, que é igual a zero para quase todos os imóveis, e o mês da venda não apresentam correlação significativa com praticamente todas as outras variáveis.
- O tamanho da área de convivência apresenta uma correlação bastante alta (¿ 0.8) com o número de cômodos.
- A área do primeiro piso tem uma correlação elevada (0.8) com a área do porão.
- Existe uma forte correlação negativa entre "Bsmt Unf SF"(espaço do porão inacabado) e "BsmtFin SF 1"(espaço do porão acabado), indicando (de maneira bastante intuitiva) que, à medida que a área do porão acabado aumenta, a área do porão inacabado diminui.

3.2 Correlação entre *features* e preço de venda

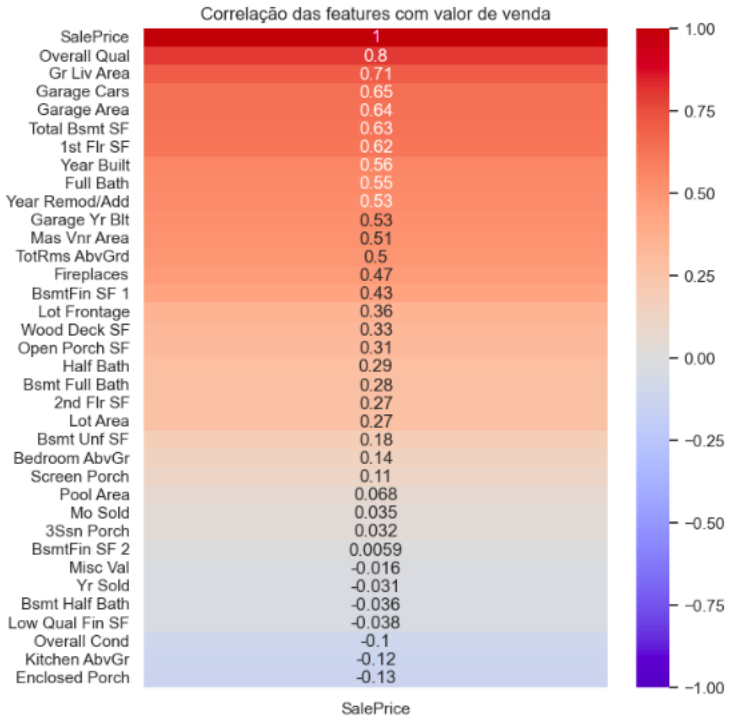

Figura 3: Correlação das *features* com preço de venda

Como mostrado na figura 3, das variáveis com maior correlação com o preço de venda, a qualidade geral do imóvel (que indica a qualidade do material e acabamento usado e é representado por uma nota de 1 a 10) é a que apresenta a correlação mais alta com o valor de venda.

Em sequência, tem-se o tamanho da área de convivência, a capacidade e área da garagem e a área do sótão com as maiores correlações.

No caso das variáveis com correlação negativa, tem-se que casas com uma área de varanda cercada maior e com mais cozinhas tem uma tendência (leve) a ter um preço menor.

No caso de varandas cercadas, conforme análise abaixo, se vê uma correlação negativa (-0.4) com a idade do imóvel. Ou seja, quanto mais antiga a casa, maior a área de varanda cercada e vice-versa.

3.3 Análise das variáveis escolhidas

3.3.1 Scatterplot das variáveis numéricas

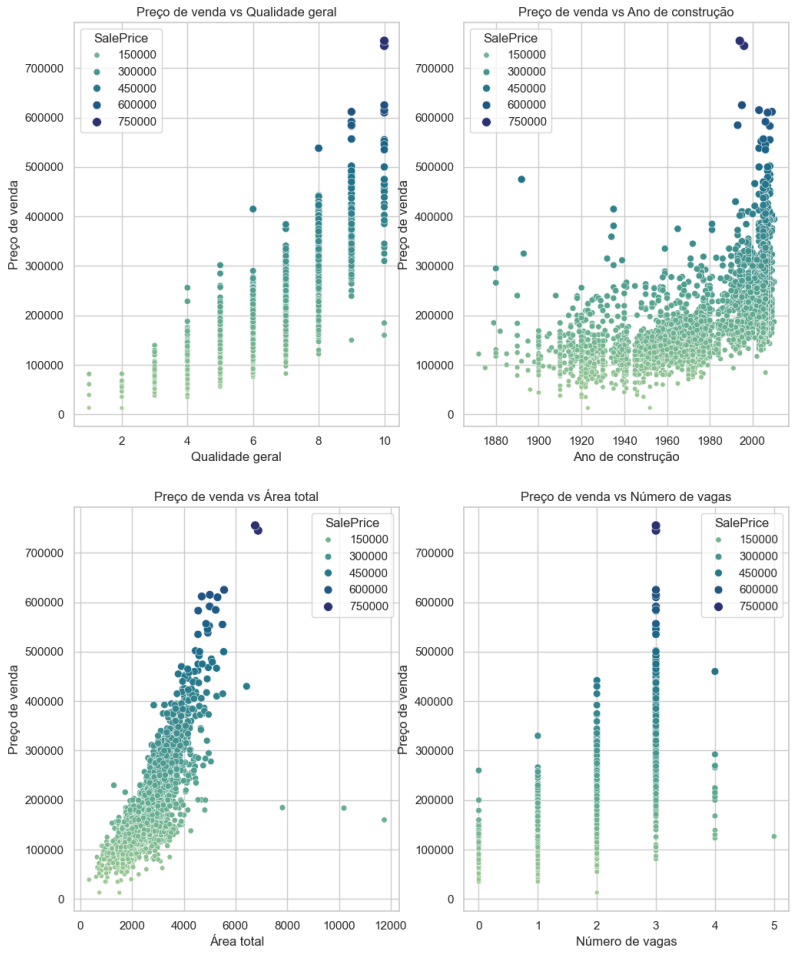

Figura 4: Scatterplots - Variáveis numéricas

O grafico 4 confirma o que o gráfico de correlação já mostra, mas dá uma ideia melhor do tipo de relação matemática entre as variáveis independentens e o preço de venda. Como pode-se ver, há uma aparente tendência linear nos 4 gráficos, mas tanto o ano de construção quanto a área total, para valores à direita, aparentam um crescimento mais acentuado, talvez exponencial.

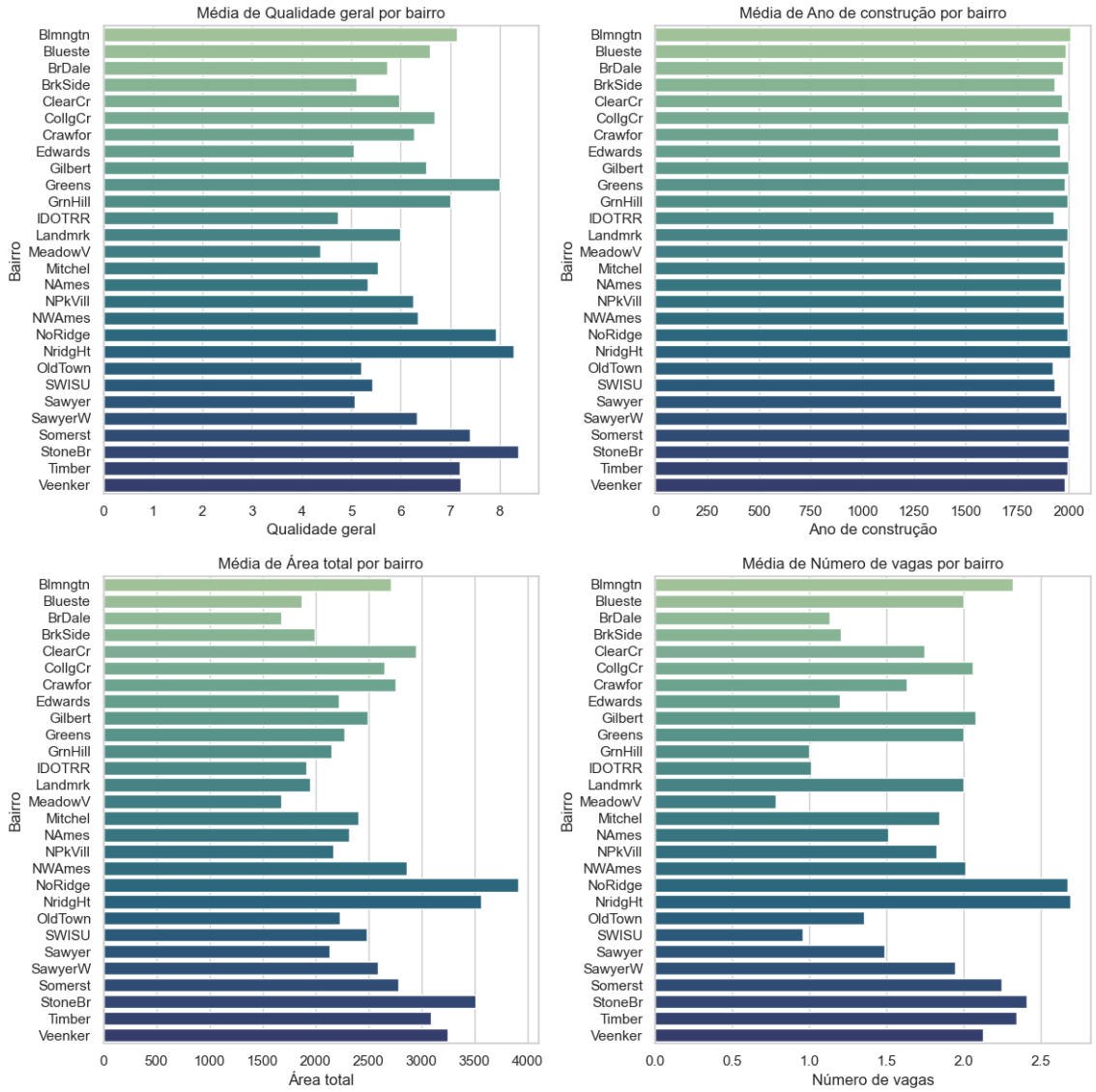

Figura 5: Distribuição da média das 4 variáveis numéricas por bairro

Já os gráficos de distribuição da média das variáveis numéricas por bairro confirmam a ideia de que há uma grande variação em relação a essas variáveis por bairro, indicando, por exemplo, interação entre a região e a qualidade geral dos imóveis.

4 Resultados da análise de pressupostos da regressão linear

Os pressupostos necessários para a realização da regressão linear foram verificados, com os seguintes resultados:

- **Normalidade**: foi verificada por meio de gráficos Q-Q plots [5] e histograma e pelo teste de Shapiro-Wilk [2]. Ambos os métodos, mostrados em 6 e 8 indicaram que os resíduos não seguem uma distribuição normal para todos os valores. Nos Q-Q plots, especialmente nos quantis superiores, as observações se desviaram da linha diagonal, sugerindo a presença de caudas pesadas e valores extremos (outliers).

Figura 6: Q-Q plots para resíduos

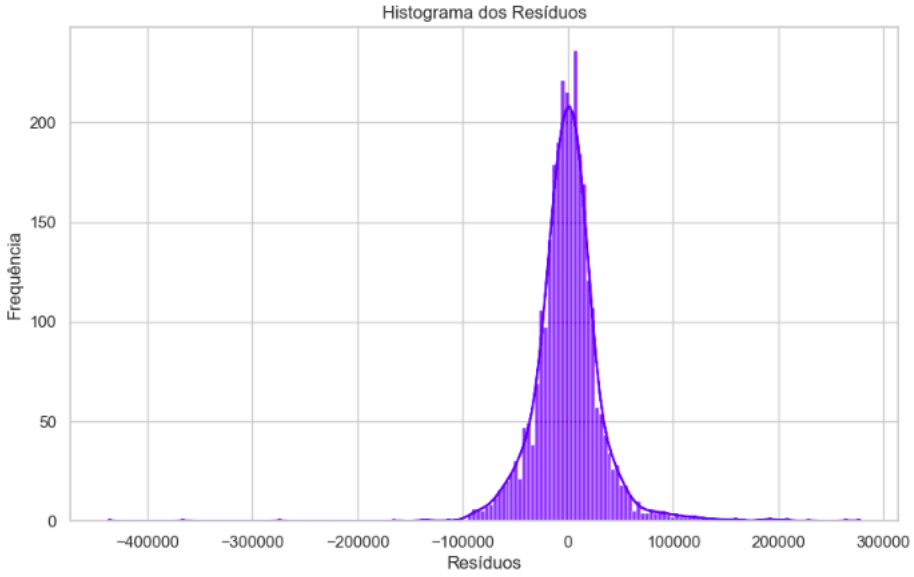

Figura 7: Histograma dos resíduos

- **Homocedasticidade**: foi testada utilizando análise visual, por meio do *scatterplot* dos resíduos, e pelo teste de White [4]. Ambos indicaram que as variâncias dos resíduos não são constantes, ou seja, os resíduos não apresentam homocedasticidade, o que sugere que os pressupostos da regressão linear não estão completamente atendidos.
- **Multicolinearidade**: foi realizada utilizando o Fator de Inflação da Variância (VIF). Verificou-se que duas variáveis, "Neighborhood_NAmes"e "Neighborhood_OldTown", apresentaram VIFs superiores ao valor limiar de 10 [3]. Como resultado, essas variáveis foram removidas da análise subsequente e da construção dos modelos, visando evitar problemas de multicolinearidade.

5 Modelos robustos

Como os pressupostos principais da regressão linear não foram atendidos, duas abordagens robustas foram testadas: um modelo de regressão linear ponderada (WLS) e uma árvore de decisão.

Os modelos foram treinados com 80% do dado original e testados com os 20% restantes.

5.1 Regressão Linear Ponderada

A regressão linear ponderada (Weighted Least Squares, WLS) foi utilizada para dar mais importância aos pontos de dados que apresentam menor variabilidade, considerando a heterocedasticidade dos resíduos. O modelo foi ajustado com pesos baseados na variância dos resíduos, permitindo que as observações com maior variabilidade (geralmente os outliers) tivessem menor impacto no modelo. Apesar de uma melhora no ajuste, o modelo ainda mostrou algumas limitações devido à presença de outliers, com um **RMSE de 37373 e R^2 de 0.82**.

Os valores preditos em relação aos valores esperados podem ser vistos em **??**.

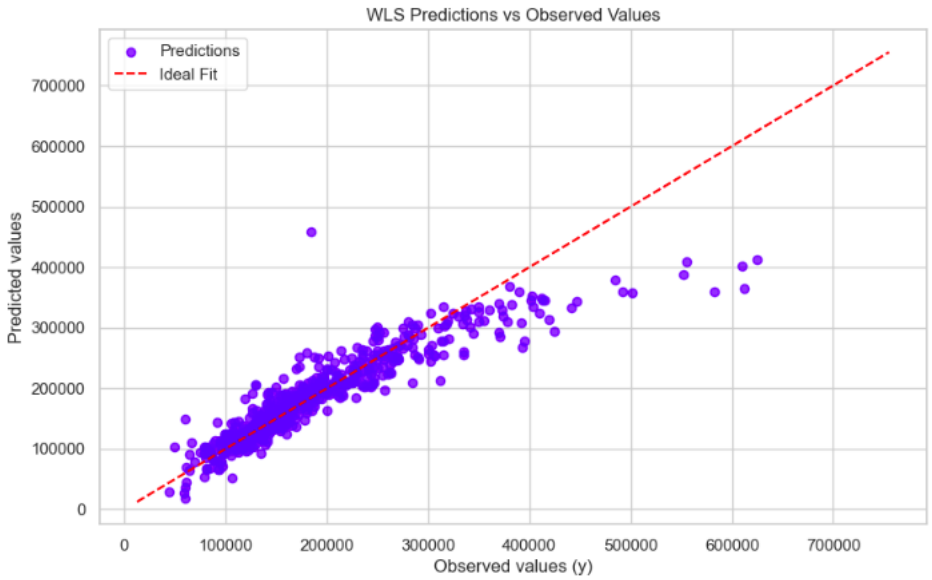

Figura 8: Histograma dos resíduos

5.1.1 Interação entre área do imóvel e qualidade geral

Para esse modelo, foi testado a criação de uma nova variável representando a interação entre área do imóvel e qualidade geral. Como uma casa de qualidade superior pode ter espaços mais luxuosos ou eficientes, tem-se que sua metragem quadrada total pode ter um impacto maior no preço. Já que a relação entre a metragem quadrada total e o preço pode ser mais forte ou mais fraca, dependendo da qualidade geral da casa, ao incluir uma interação entre essas duas variáveis, é possível testar a hipótese de que o efeito da metragem quadrada total sobre o preço da casa não é constante, mas varia de acordo com a qualidade da casa.

Após essa inclusão, o modelo apresentou as seguintes métricas: **RMSE de 35874.9 e R^2 de 0.84**, melhor do que sua versão anterior.

Avaliando o p-valor e o coeficiente gerado para essa nova variável no sumário do modelo (marcado em azul na figura 9), tem-se um coeficiente com valor significativo e um p-valor extremamente baixo, o que indica que o termo de interação é estatisticamente relevante.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.355e+05	4.73e+04	-11.324	0.000	-6.28e+05	-4.43e+05
Overall Qual	4047.1771	669.747	6.043	0.000	2733.810	5360.545
Year Built	292.6420	24.226	12.080	0.000	245.136	340.148
Total SF	4.5176	1.979	2.282	0.023	0.636	8.399
Garage Cars	5661.7204	684.910	8.266	0.000	4318.619	7004.822
Neighborhood_Blueste	-1.863e+04	7612.808	-2.447	0.014	-3.36e+04	-3701.645
Neighborhood_BrDale	-2.516e+04	3700.826	-6.799	0.000	-3.24e+04	-1.79e+04
Neighborhood_BrkSide	386.1113	1722.887	0.224	0.823	-2992.453	3764.676
Neighborhood_ClearCr	2.51e+04	4864.498	5.160	0.000	1.56e+04	3.46e+04
Neighborhood_CollgCr	4154.7653	2199.704	1.889	0.059	-158.833	8468.363
Neighborhood_Crawfor	2.367e+04	2914.088	8.123	0.000	1.8e+04	2.94e+04
Neighborhood_Edwards	3075.2205	1431.913	2.148	0.032	267.254	5883.187
Neighborhood_Gilbert	7197.9282	2707.462	2.659	0.008	1888.622	1.25e+04
Neighborhood_Greens	-1250.4688	1.19e+04	-0.105	0.917	-2.47e+04	2.22e+04
Neighborhood_GrnHill	1.007e+05	1.51e+04	6.690	0.000	7.12e+04	1.3e+05
Neighborhood_IDOTRR	-8866.5397	1631.936	-5.433	0.000	-1.21e+04	-5666.329
Neighborhood_Landmrk	-1.692e+04	2.1e+04	-0.805	0.421	-5.82e+04	2.43e+04
Neighborhood_MeadowV	-1.329e+04	2618.668	-5.075	0.000	-1.84e+04	-8154.499
Neighborhood_Mitchel	6415.0892	2197.052	2.920	0.004	2106.693	1.07e+04
Neighborhood_NPkVill	-1.823e+04	6250.418	-2.916	0.004	-3.05e+04	-5969.889
Neighborhood_NWAmes	425.5856	3069.096	0.139	0.890	-5592.882	6444.053
Neighborhood_NoRidge	3.892e+04	7282.982	5.344	0.000	2.46e+04	5.32e+04
Neighborhood_NridgHt	2.829e+04	4850.606	5.832	0.000	1.88e+04	3.78e+04
Neighborhood_SWISU	-309.9233	4070.525	-0.076	0.939	-8292.184	7672.337
Neighborhood_Sawyer	4167.0748	1739.441	2.396	0.017	756.048	7578.101
Neighborhood_SawyerW	-2758.3515	2782.316	-0.991	0.322	-8214.446	2697.743
Neighborhood_Somerst	7557.4125	3220.642	2.347	0.019	1241.766	1.39e+04
Neighborhood_StoneBr	2.463e+04	8095.881	3.043	0.002	8756.330	4.05e+04
Neighborhood_Timber	1.489e+04	4731.661	3.147	0.002	5610.939	2.42e+04
Neighborhood_Veenker	2.421e+04	8076.024	2.998	0.003	8375.060	4e+04
OverallQual_TotalSF_interaction	5.2783	0.360	14.681	0.000	4.573	5.983

Omnibus:	434.958	Durbin-Watson:	1.953
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6460.060
Skew:	-0.424	Prob(JB):	0.00
Kurtosis:	11.088	Cond. No.	1.31e+06

Figura 9: Sumário do modelo de WLS com interação de *features*

5.2 Árvore de Decisão

Como alternativa robusta, foi testada uma árvore de decisão regressora. Esse modelo não depende dos pressupostos de linearidade, normalidade ou homocedasticidade, e pode capturar relações não-lineares entre as variáveis preditoras e o preço de venda. A árvore de decisão teve um resultado pior (**RMSE de 42244.6 e R^2 de 0.77**) em relação ao WLS.

As predições podem ser vistas em 10.

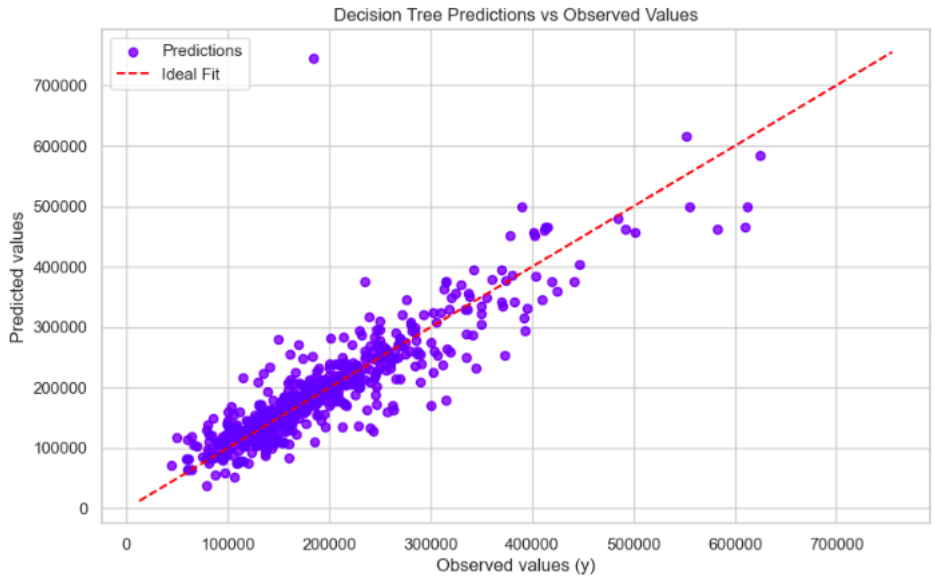

Figura 10: Predições da árvore de decisão

6 Conclusão

As análises indicam que as características com maior impacto no preçoo de venda incluem *Overall Qual*, *Total SF*, *Garage Cars* e *Year Built*. A qualidade geral e a *Total SF* destacam-se como os principais preditores devido à sua forte correlação com o preç.

O uso de modelos robustos mostrou-se justificado pela presença de heterocedasticidade e *outliers*. Apesar disso, a regressão linear ponderada (WLS) apresentou um desempenho superior em relação à *árvore de decisãao*, sugerindo que, com o tratamento adequado de *outliers* e variáveis, o modelo linear ainda é uma abordagem eficaz. Trabalhos futuros podem focar em técnicas como *Gradient Boosting* ou *Random Forests* para capturar relações não lineares e melhorar a capacidade preditiva.

Mais espeficamente, a performance do modelo WLS foi avaliada utilizando o Root Mean Square Error (RMSE) no conjunto de teste, obtendo um valor de 37.373. Este valor indica o erro médio na previsão dos preços de venda. Em outras palavras, o preço previsto pelo modelo possui, em média, um desvio de aproximadamente 37 mil dólares em relação ao preço real.

Nesse sentido, embora o modelo WLS forneça boas previsões em geral, ele não é perfeito para todos os contextos. Para uma imobiliária, o modelo seria útil em cenários de avaliação rápida e para imóveis típicos. No entanto, para previsões mais precisas e robustas, especialmente em casos complexos ou com presença de outliers, combinar o WLS com métodos não lineares e técnicas de regularização seria a estratégia mais eficaz.

A inclusão de mais variáveis independentes na modelagem, apesar de torná-la mais complexa, poderia gerar resultados mais consistentes e precisos também.

Referências

[1] Iowa Ames. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3):161–169, 2011.

[2] Universidade Estadual de Londrina. Teste de shapiro-wilk. `https://www.uel.br/projetos/experimental/pages/arquivos/Shapiro.html`. Accessed: 2024-12-08.

[3] Lüdecke, Daniel and Ben-Shachar, Mattan S. and Patil, Indrajeet and Waggoner, Philip and Makowski, Dominique. *check_collinearity function*. easystats, 2024. Accessed: 2024-06-15.

[4] Alexandre Gori Maia. Heterocedasticidade, 2023. Accessed: 2024-15-12.

[5] Statology. How to use q-q plots to check normality. `https://www.statology.org/q-q-plot-normality/`. Accessed: 2024-12-08.

[6] Jared Wilber. Linear regression. `https://mlu-explain.github.io/linear-regression/`. Accessed: 2024-15-12.