



Técnicas de Amostragem e Regressão Linear
Atividade Final

Gabriel Fratucci dos Reis
2025

Descrição da Atividade Final

Realizar um projeto utilizando um dos modelos apresentados na disciplina. O Trabalho pode ser feito em grupo com número máximo de 2 alunos por grupo.

- 1) Acessar o <https://www.kaggle.com/>
- 2) Selecionar a base de escolha e que se adeque ao modelo.
- 3) Contextualizar o problema a ser resolvido.
- 4) processamento de dados
- 5) Análise de variáveis - análise descritiva
- 6) análise de correlação
- 7) validação de pressupostos
- 8) rodar o modelo
- 9) interpretação da saída (métricas de avaliação e coeficientes)

Entregar o notebook (em python ou r com as saídas comentadas com a sua interpretação) + a base de dados utilizada.

Contextualização do Projeto

O objetivo deste projeto é o treinamento e a aplicação de um modelo, usando regressão linear múltipla, sobre a base de dados KC House Data extraída no <https://www.kaggle.com/>.

O conjunto de dados consiste em preços de imóveis no Condado de King, uma área no estado americano de Washington. Esses dados também abrangem Seattle. A base consistiu em 21 variáveis e 21.613 observações.

O objetivo do modelo é prever os preços dos imóveis com base em suas variáveis, para isso foram observados:

- Limpeza e tratamento dos dados da base
- Coleta da amostra da base e verificação das correlações
- Resultados do modelo

Desenvolvimento do Projeto

Limpeza e Tratamento de Dados

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built
count	2.161300e+04	2.161300e+04	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	4.580302e+09	5.400881e+05	3.370842	2.114757	2079.899736	15.10697e+04	1.494309	0.007542	0.234303	3.409430	7.656873	1788.390691	291.509045	1971.005
std	2.876566e+09	3.671272e+05	0.930062	0.770163	918.440897	4.142051e-04	0.539989	0.086517	0.766318	0.650743	1.175459	828.090978	442.575043	29.371
min	1.000102e+06	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	1.000000	290.000000	0.000000	1900.000
25%	2.123049e+09	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000	7.000000	1190.000000	0.000000	1951.000
50%	3.904930e+09	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000	3.000000	7.000000	1560.000000	0.000000	1975.000
75%	7.308900e+09	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000	4.000000	8.000000	2210.000000	560.000000	1997.000
max	9.900000e+09	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000	5.000000	13.000000	9410.000000	4820.000000	2015.000

Remoção de outliers, temos terrenos com 33 quartos, em uma média de 3, na variável bedrooms

Alguns terrenos não possuem quartos nem banheiros, tornando-se um erro, nas variáveis, bedrooms e bathrooms

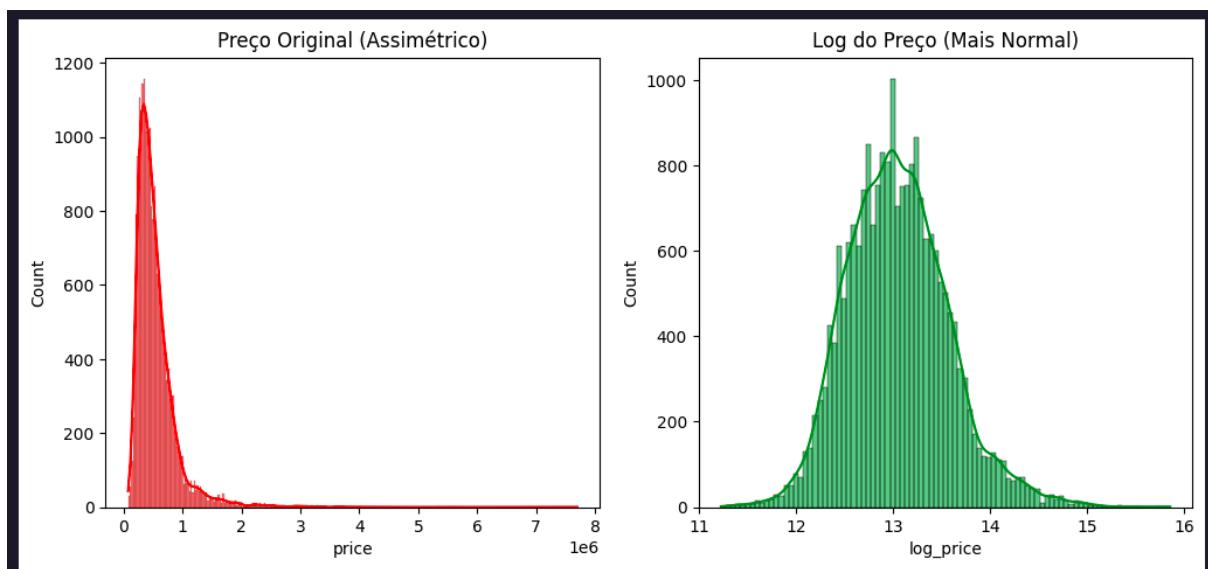
Além disso, temos terrenos absurdamente grandes, o que pode indicar um erro, na variável sqft_lot

Remoção da var id (é irrelevante)

Muito mais importante que a data de construção da casa, é o intervalo de tempo entre a data de construção e a data atual

Indicação de reforma na casa para a variável was_renovated

Normalização da variável price, que possui uma assimetria à direita nos dados, ao aplicar a transformação logarítmica, os dados se aproximam da distribuição normal



Coleta da Amostra

Pelo tamanho do da base, com 21000 registros aproximadamente, nós poderíamos usar a base completa para treinar o modelo, mas para fins de estudo, vamos realizar uma amostra estratificada sobre a variável grade.

A variável grade é a avaliação ou "nota" da casa. Como algumas notas possuem poucos dados, como por exemplo, 12, 4 e 13, se realizarmos uma amostra aleatória simples, pode-se ocorrer o viés de seleção

Por que escolher a variável grade e não outras como o zipcode ou bedrooms? Pois ela possui a melhor correlação com a nossa variável preço. Então é muito importante que o modelo aprenda sobre essa variável.

Além disso consideramos a remoção de algumas variáveis após avaliar a matriz de correlação de Pearson:

sqft_above: Altamente correlacionada com sqft_living (0.88) Área construída acima do nível do solo (exclui o porão).

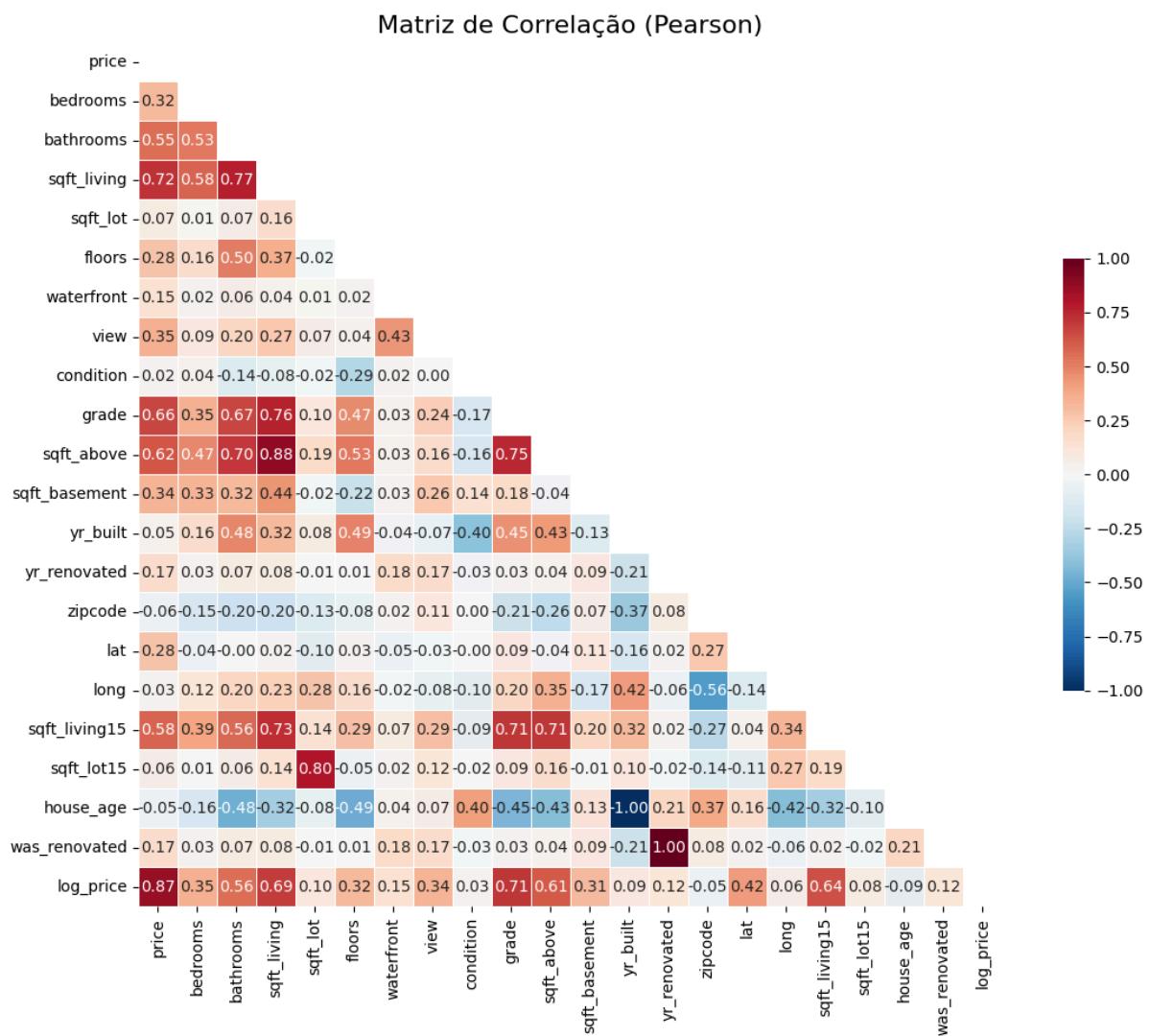
sqft_lot15: Altamente correlacionada com sqft_lot (0.80) Área do Terreno dos 15 vizinhos mais próximos

yr_built: Redundante com house_age (-1.0)

sqft_living15: Redundante com sqft_living (0.73) Área Útil ou Área Construída interna

price: Alvo original (já temos log_price)

zipcode: Variável nominal tratada como numérica (ruído)



Resultados do Modelo

OLS Regression Results												
Dep. Variable:	log_price	R-squared:	0.763									
Model:	OLS	Adj. R-squared:	0.761									
Method:	Least Squares	F-statistic:	319.2									
Date:	Fri, 12 Dec 2025	Prob (F-statistic):	0.00									
Time:	21:44:02	Log-Likelihood:	-80.968									
No. Observations:	1500	AIC:	193.9									
Df Residuals:	1484	BIC:	278.9									
Df Model:	15											
Covariance Type:	nonrobust											
	coef	std err	t	P> t	[0.025	0.975]						
const	-53.8218	7.147	-7.530	0.000	-67.842	-39.802						
bedrooms	-0.0005	0.010	-0.057	0.955	-0.019	0.018						
bathrooms	0.0432	0.016	2.741	0.006	0.012	0.074						
sqft_living	0.0002	1.59e-05	12.085	0.000	0.000	0.000						
sqft_lot	6.265e-07	2.1e-07	2.980	0.003	2.14e-07	1.04e-06						
floors	0.0514	0.017	2.979	0.003	0.018	0.085						
waterfront	0.4132	0.075	5.479	0.000	0.265	0.561						
view	0.0656	0.010	6.725	0.000	0.046	0.085						
condition	0.0742	0.012	6.324	0.000	0.051	0.097						
grade	0.1796	0.010	18.162	0.000	0.160	0.199						
sqft_basement	-3.427e-05	2.08e-05	-1.650	0.099	-7.5e-05	6.48e-06						
...												
Notes:												
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.												
[2] The condition number is large, 3.96e+07. This might indicate that there are strong multicollinearity or other numerical problems.												

Ao rodarmos o primeiro treinamento, apesar do $R^2 = 76.3\%$, verificamos uma multicolinearidade alta, isso se dá pela mistura de números grandes com números pequenos, para corrigir isso, podemos usar o StandardScaler, e ao padronizar o modelo obtemos os seguintes resultados

StandardScaler -> Ao padronizar os dados, podemos verificar que o grade tem a maior importância no modelo

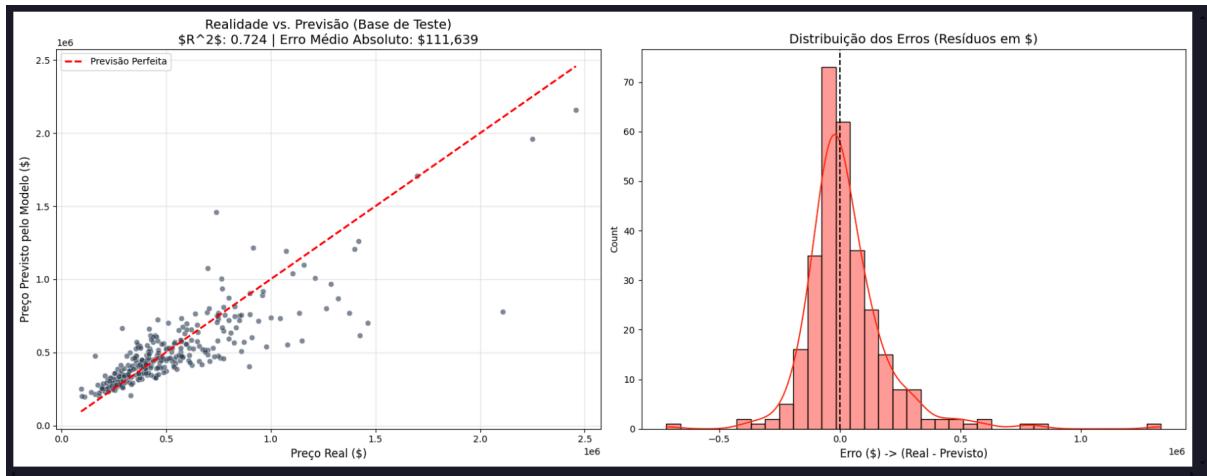
Teste de Hipótese -> Podemos verificar que a variável bathrooms é irrelevante para o modelo, possivelmente obteremos um modelo melhor ao retirá-la, o que reduzirá o AIC e o BIC

```

OLS Regression Results
=====
Dep. Variable: log_price R-squared: 0.761
Model: OLS Adj. R-squared: 0.759
Method: Least Squares F-statistic: 337.5
Date: Fri, 12 Dec 2025 Prob (F-statistic): 0.00
Time: 21:44:02 Log-Likelihood: -89.017
No. Observations: 1500 AIC: 208.0
Df Residuals: 1485 BIC: 287.7
Df Model: 14
Covariance Type: nonrobust
=====
            coef    std err      t      P>|t|      [0.025]     [0.975]
-----
const      13.0560   0.007  1959.458   0.000    13.043    13.069
bedrooms    0.0016   0.009    0.189   0.850   -0.015    0.019
bathrooms   0.0333   0.012    2.678   0.007    0.009    0.058
sqft_living  0.1787   0.015   11.897   0.000    0.149    0.208
sqft_lot     0.0209   0.007    2.931   0.003    0.007    0.035
floors       0.0284   0.009    3.048   0.002    0.010    0.047
waterfront    0.0390   0.008    5.183   0.000    0.024    0.054
view         0.0552   0.008    6.841   0.000    0.039    0.071
condition    0.0452   0.008    6.011   0.000    0.030    0.060
grade        0.2121   0.012   18.177   0.000    0.189    0.235
sqft_basement -0.0135   0.009   -1.459   0.145   -0.032    0.005
...
=====
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

Resultado Final:



Podemos verificar os seguintes pontos:

R2 -> O modelo explica 76% das variações do modelo.

AIC de 208.0 -> Números baixos indicam um bom equilíbrio entre erro e simplicidade.

O Coeficiente é 0.0016 e o P-valor ($P>|t|$) é 0.850. Isso prova estatisticamente que a quantidade de quartos é irrelevante para o preço final neste modelo.

As variáveis grade (0.2121) e sqft_living (0.1787) têm os maiores coeficientes e P-valor 0.000. Eles têm maior influência no modelo.

O histograma de resíduos confirma que o modelo não é viciado.

Ponto importante:

O modelo subestima casas com valores altos.

Considerando esses pontos, podemos afirmar que temos um modelo preditivo funcional.