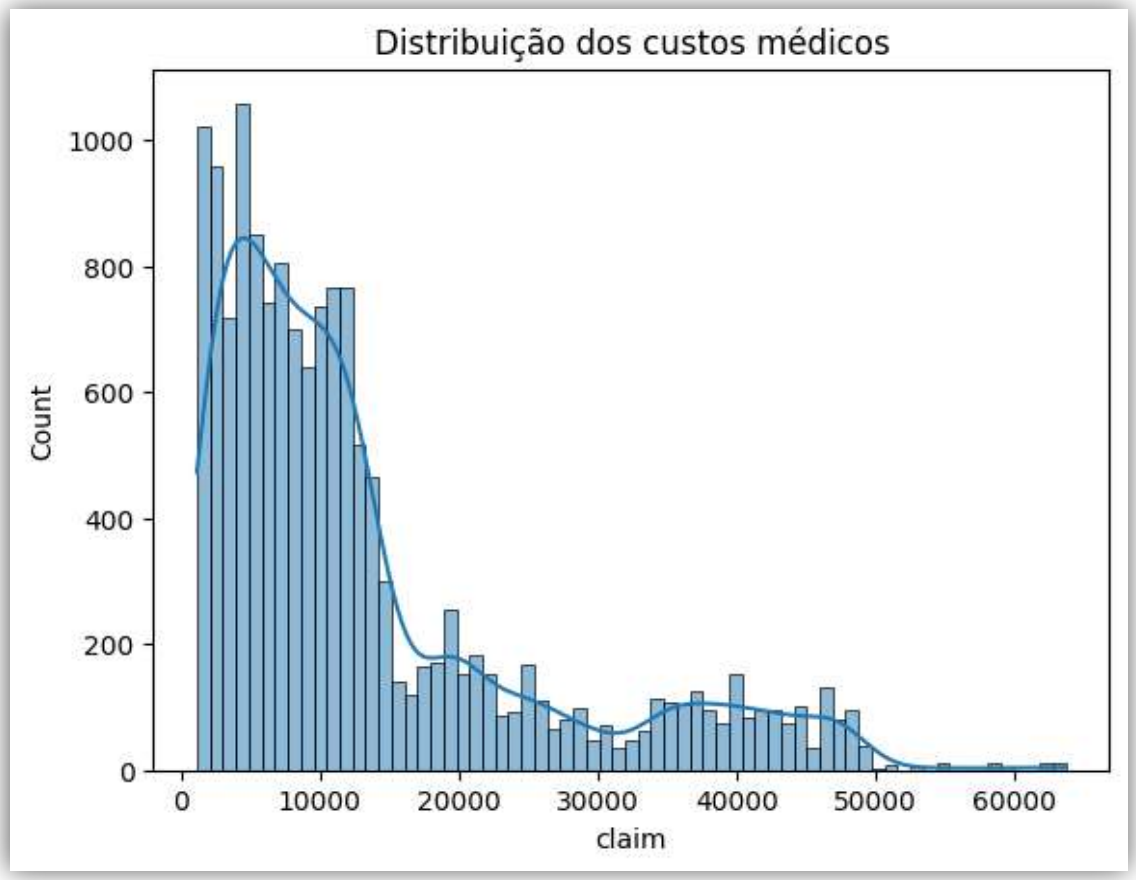
Relatório de Análise de Custos Médicos

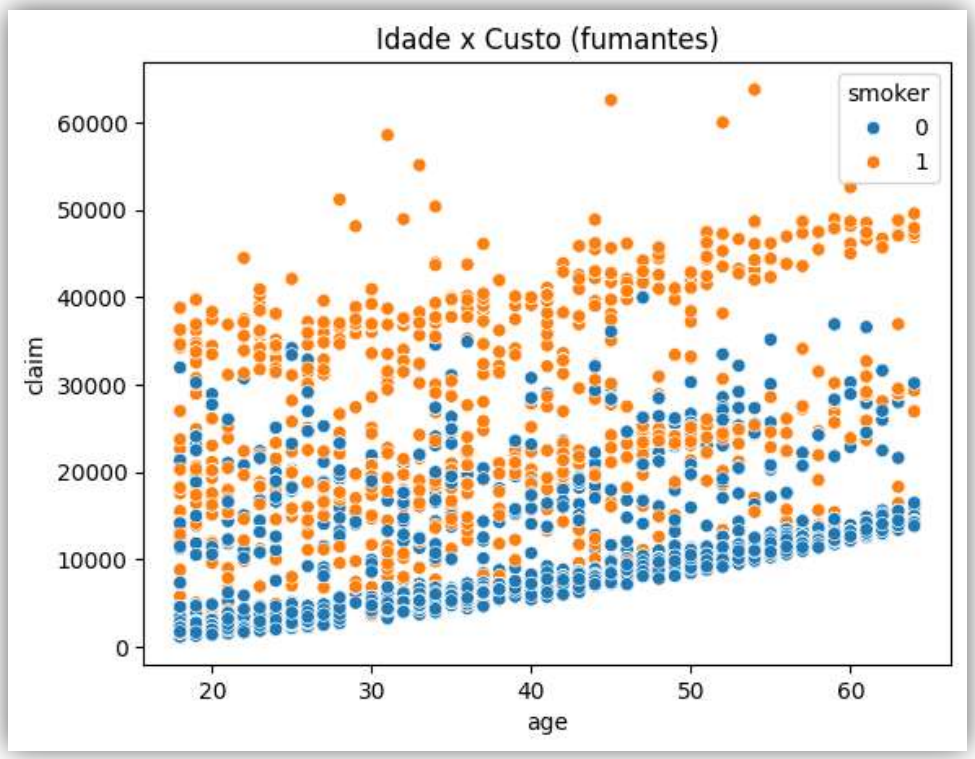
Os 10 primeiros registros da Base original:

age	sex	weight	bmi	hereditary_diseases	no_of_dependents	smoker	city	bloodpressure	diabetes	regular_ex	claim
60.0	male	64	24.3	NoDisease	1	0	NewYork	72	0	0	13.112,60
49.0	female	75	22.6	NoDisease	1	0	Boston	78	1	1	9.567,00
32.0	female	64	17.8	Epilepsy	2	1	Phildelphia	88	1	1	32.734,20
61.0	female	53	36.4	NoDisease	1	1	Pittsburg	72	1	0	48.517,60
19.0	female	50	20.6	NoDisease	0	0	Buffalo	82	1	0	1.731,70
42.0	female	89	37.9	NoDisease	0	0	AtlanticCity	78	0	0	6.474,00
18.0	male	59	23.8	NoDisease	0	0	Portland	64	0	0	1.705,60
21.0	male	52	26.8	NoDisease	0	0	Cambridge	74	1	0	1.534,30
63.0	male	55		NoDisease	0	0	Hartford	70	1	0	13.390,60
40.0	female	69	29.6	NoDisease	0	0	Springfield	64	1	1	5.910,90

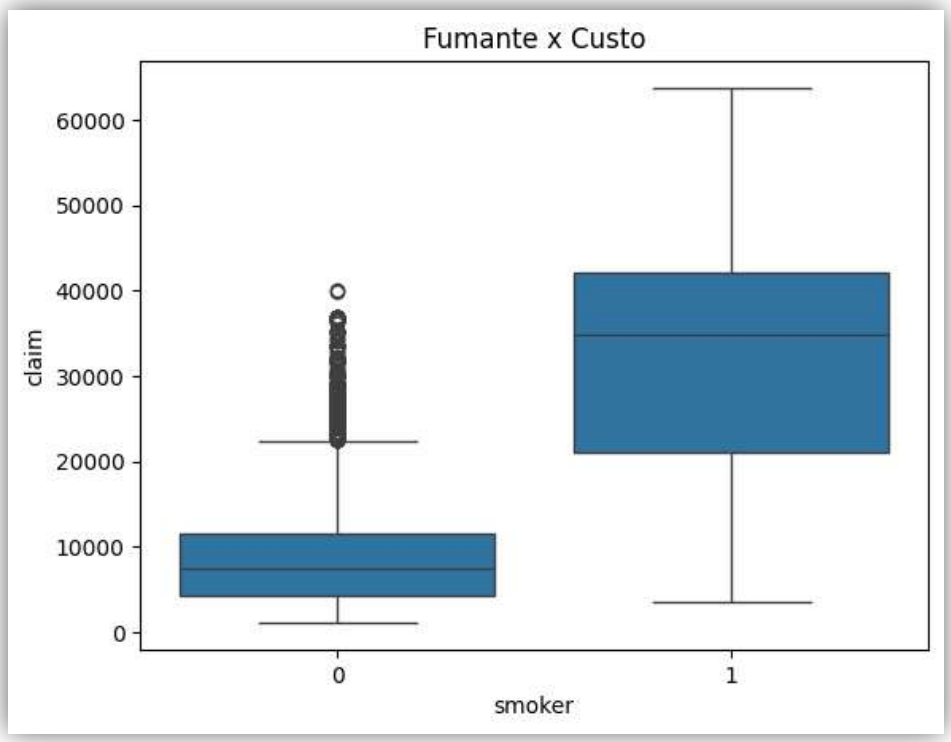
Análises Exploratórias



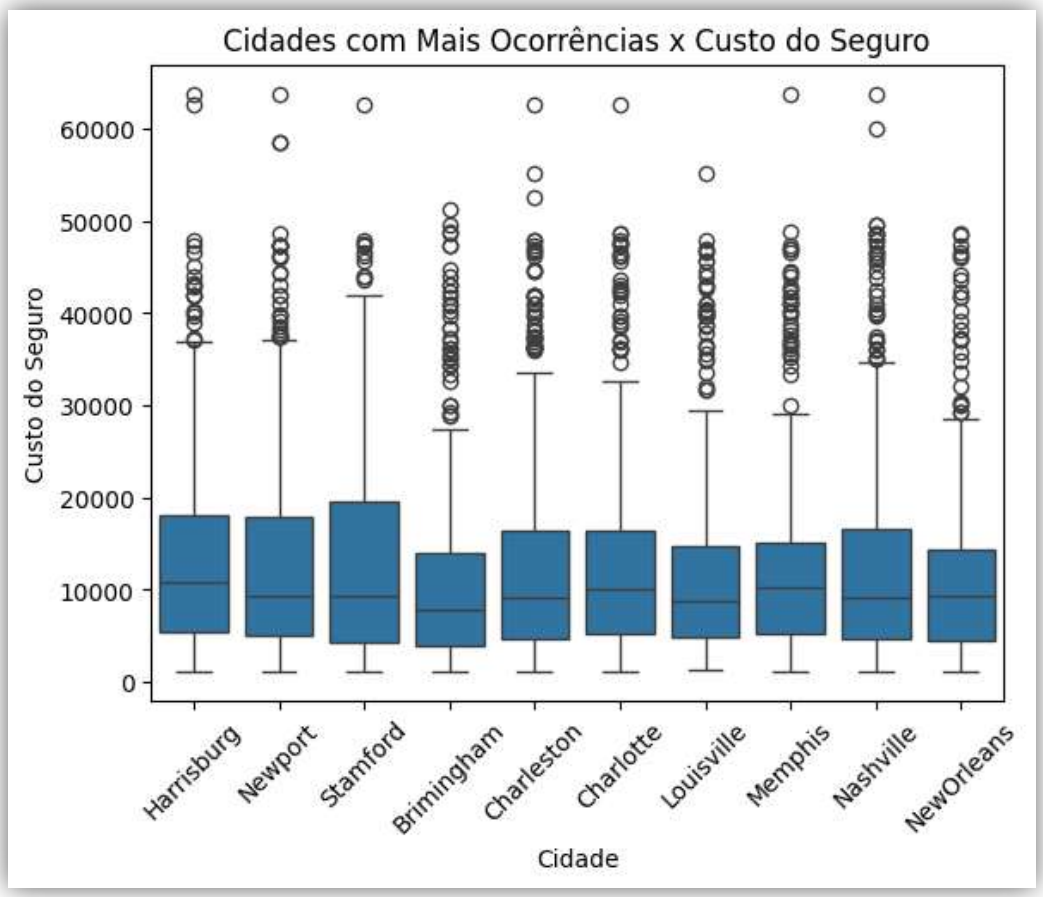
O histograma da Distribuição dos Custos Médicos revela que a maioria dos indivíduos apresenta custos médicos mais baixos, com uma diminuição progressiva na frequência à medida que os custos aumentam. A assimetria à direita da distribuição sugere a presença de alguns casos com custos médicos significativamente elevados.



O gráfico de dispersão Idade x Custo (fumantes) ilustra a relação entre a idade e os custos médicos exclusivamente para o grupo de fumantes, permitindo a análise de tendências específicas dessa população.



A análise comparativa entre fumantes e não fumantes demonstra claramente que os fumantes incorrem em custos médicos consideravelmente superiores. Adicionalmente, a maior dispersão dos dados entre os fumantes aponta para uma maior variabilidade nos custos de saúde dentro desse grupo.



O boxplot Cidades com Mais Ocorrências x Custo do Seguro compara os custos do seguro nas 10 cidades com maior número de ocorrências. A análise visual permite identificar variações nos custos medianos e na dispersão dos dados entre as cidades, além da presença de outliers que indicam custos atípicos.

Amostra de dados depois do tratamento

age	weight	bmi	no_of_dependents	smoker	bloodpressure	diabetes	regular_ex	parent	sex_male	faixa_etaria
60.0	64	24.3	1	0	72	0	0	True	True	56-65
49.0	75	22.6	1	0	78	1	1	True	False	46-55
32.0	64	17.8	2	1	88	1	1	True	False	26-35
61.0	53	36.4	1	1	72	1	0	True	False	56-65
19.0	50	20.6	0	0	82	1	0	False	False	18-25
42.0	89	37.9	0	0	78	0	0	False	False	36-45
18.0	59	23.8	0	0	64	0	0	False	True	18-25
21.0	52	26.8	0	0	74	1	0	False	True	18-25
40.0	69	29.6	0	0	64	1	1	False	False	36-45
51.0	50	33.0	0	1	0	1	0	False	False	46-55

Resultados do treinamento inicial

R^2: 0.9609072515534081

MAE: 482.1655767914533

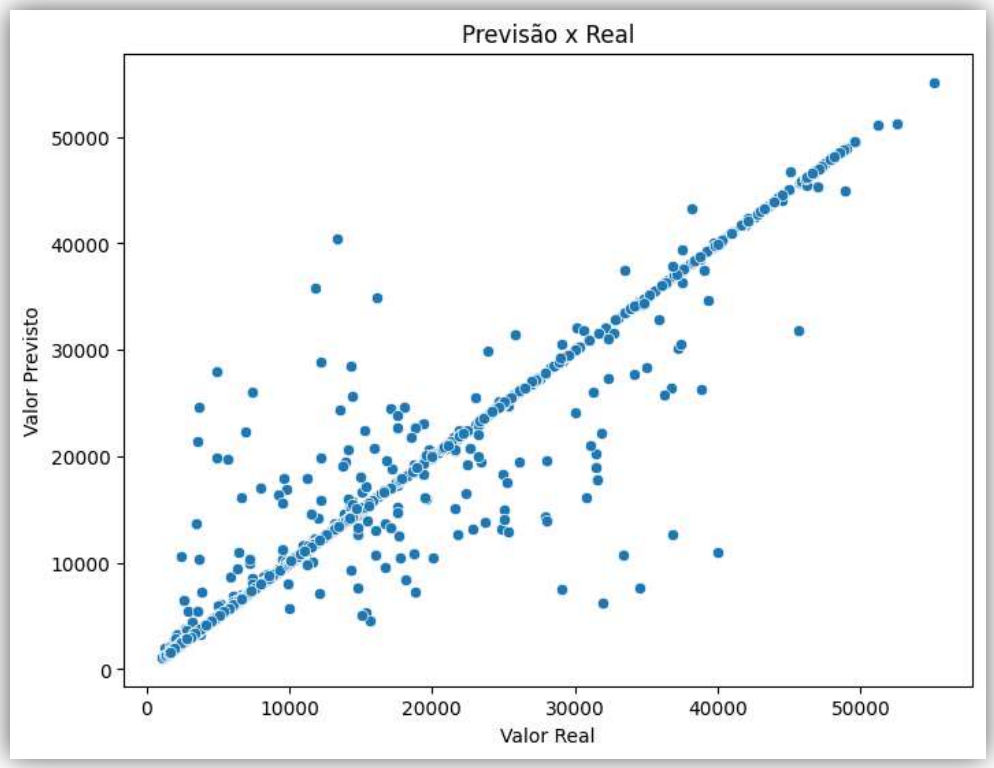
RMSE: 2377.8222727712287

OLS Regression Results

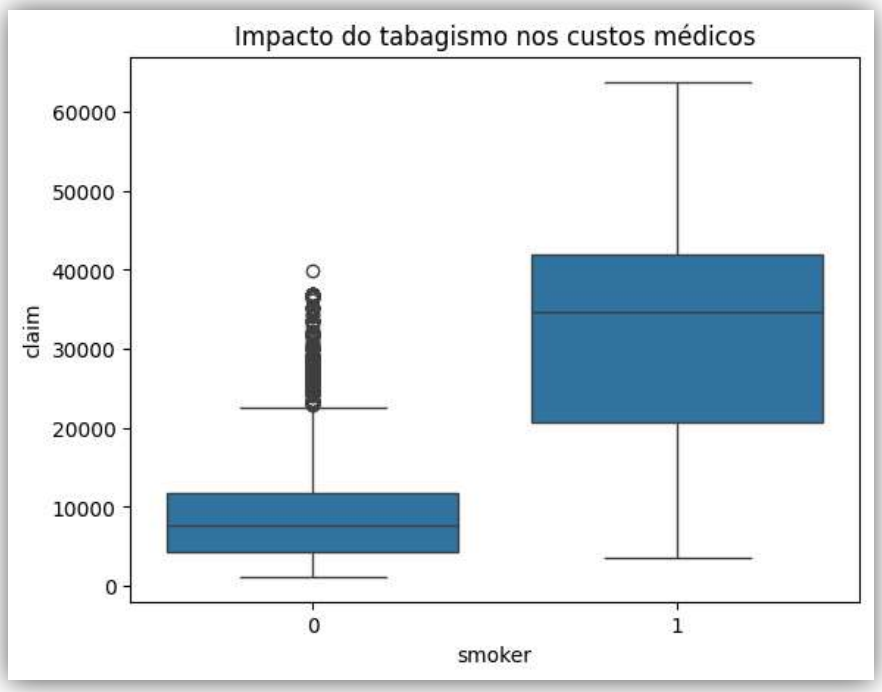
Dep. Variable:	claim	R-squared:	0.716
Model:	OLS	Adj. R-squared:	0.716
Method:	Least Squares	F-statistic:	2833.
Date:	Tue, 03 Jun 2025	Prob (F-statistic):	0.00
Time:	05:22:38	Log-Likelihood:	-1.0303e+05
No. Observations:	10113	AIC:	2.061e+05
Df Residuals:	10103	BIC:	2.062e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.354e+04	90.044	150.415	0.000	1.34e+04	1.37e+04
age	3734.0438	67.735	55.127	0.000	3601.269	3866.818
weight	-649.5132	70.003	-9.278	0.000	-786.733	-512.293
bmi	1694.6359	67.036	25.279	0.000	1563.231	1826.040
no_of_dependents	502.7146	64.796	7.758	0.000	375.701	629.728
smoker	9208.5277	64.382	143.030	0.000	9082.326	9334.729
bloodpressure	220.5152	64.139	3.438	0.001	94.791	346.240
diabetes	641.3938	64.587	9.931	0.000	514.790	767.998
regular_ex	-485.8539	64.204	-7.567	0.000	-611.706	-360.001
sex_male	-304.6471	130.555	-2.333	0.020	-560.560	-48.734

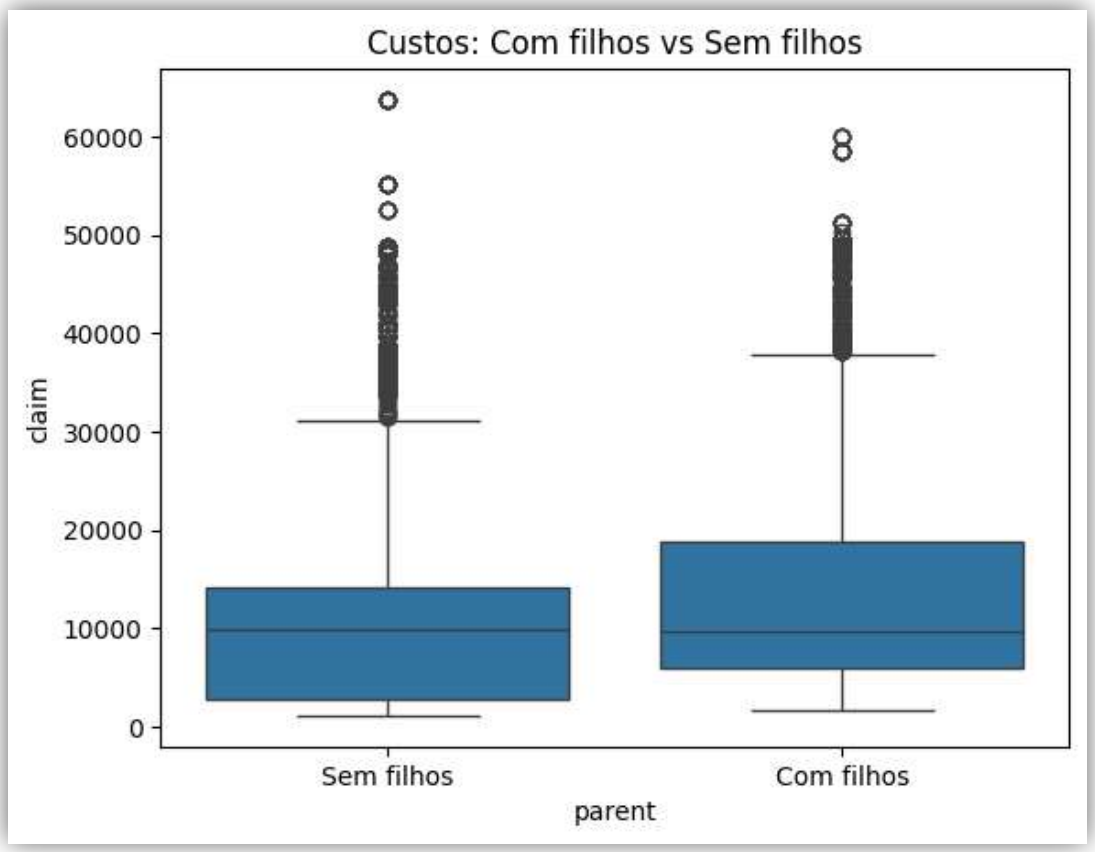
Omnibus:	1751.448	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4197.697
Skew:	0.978	Prob(JB):	0.00
Kurtosis:	5.477	Cond. No.	2.91



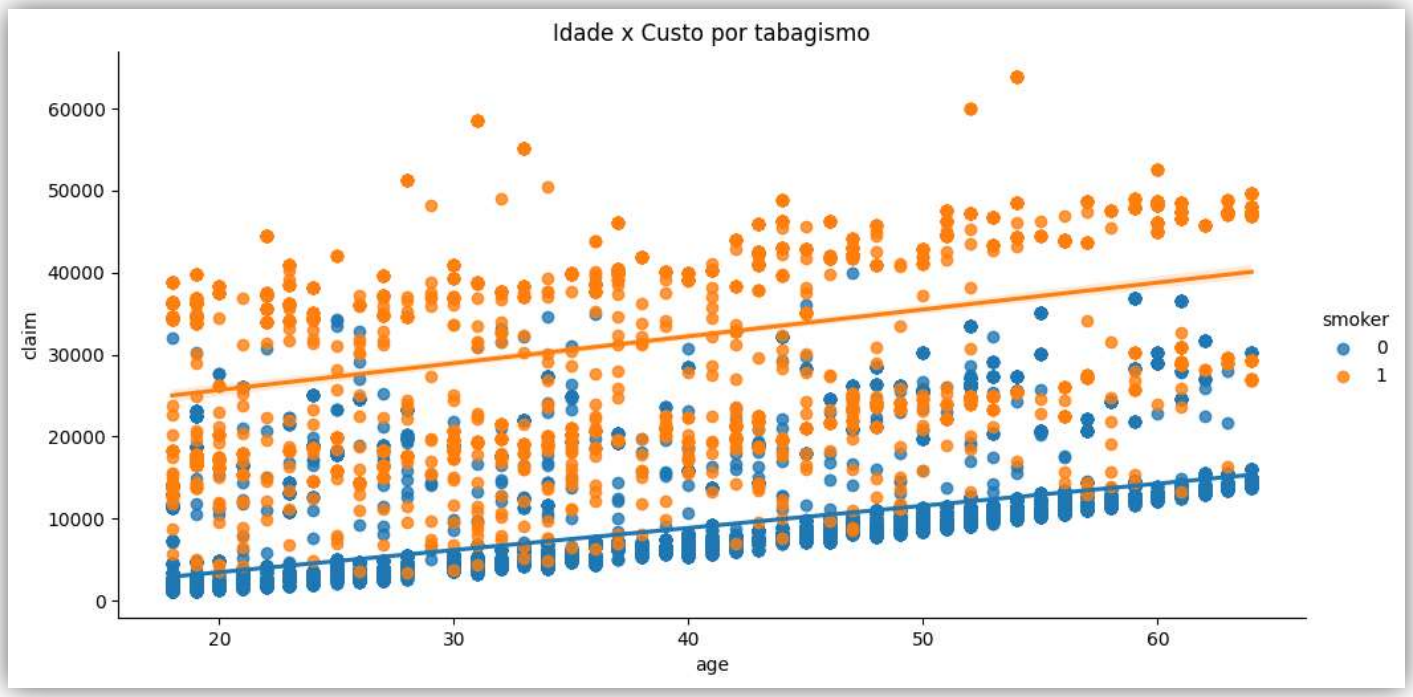
O gráfico Previsão x Real compara os valores previstos pelo modelo com os custos médicos reais. A proximidade dos pontos a uma linha diagonal indica a acurácia das previsões do modelo, enquanto desvios dessa linha revelam discrepâncias entre os valores previstos e reais.



A análise comparativa entre fumantes e não fumantes demonstra claramente que os fumantes incorrem em custos médicos consideravelmente superiores. Adicionalmente, a maior dispersão dos dados entre os fumantes aponta para uma maior variabilidade nos custos de saúde dentro desse grupo.



O boxplot Custos: Com filhos vs Sem filhos compara os custos médicos entre indivíduos com e sem filhos, analisando o impacto da presença de dependentes nos gastos com saúde.



O gráfico Idade x Custo por tabagismo examina a influência da idade nos custos médicos, diferenciando entre fumantes e não fumantes. As linhas de tendência auxiliam na visualização da progressão dos custos em relação à idade para cada grupo.

Amostra dos dados depois da remoção de outliers

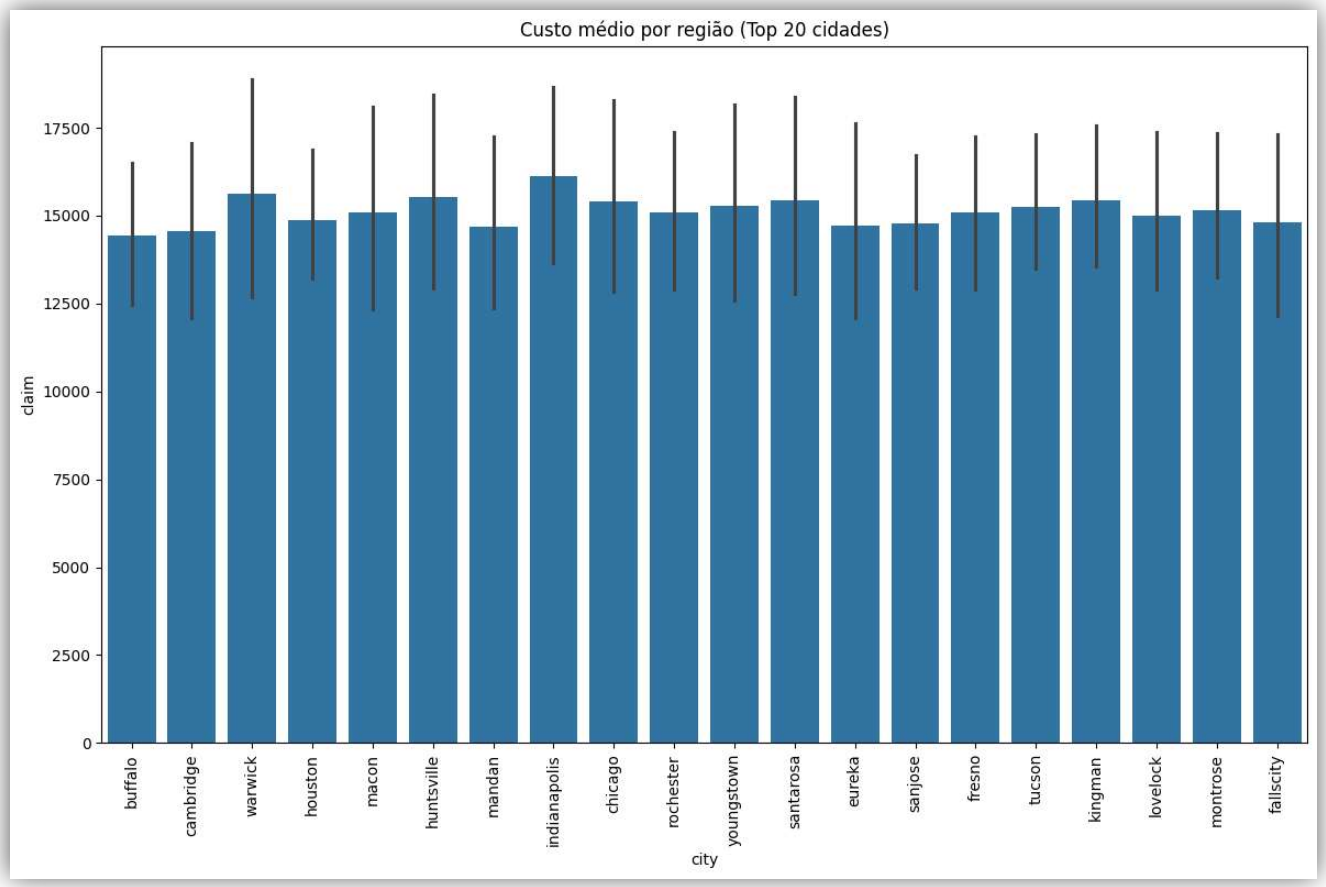
age	weight	bmi	no_of_dependents	smoker	bloodpressure	diabetes	regular_ex	claim	parent	sex_male	city	hereditary_diseases
60.0	64	24.3	1	0	72	0	0	13112.6	True	True	newyork	NoDisease
49.0	75	22.6	1	0	78	1	1	9567.0	True	False	boston	NoDisease
32.0	64	17.8	2	1	88	1	1	32734.2	True	False	phildelphia	Epilepsy
61.0	53	36.4	1	1	72	1	0	48517.6	True	False	pittsburg	NoDisease
19.0	50	20.6	0	0	82	1	0	1731.7	False	False	buffalo	NoDisease
42.0	89	37.9	0	0	0	0	0	6474.0	False	False	atlanticcity	NoDisease
18.0	59	23.8	0	0	64	0	0	1705.6	False	True	portland	NoDisease
21.0	52	26.8	0	0	74	1	0	1534.3	False	True	cambridge	NoDisease
40.0	69	29.6	0	0	64	1	1	5910.9	False	False	springfield	NoDisease
51.0	50	33.0	0	1	0	1	0	44400.4	False	False	syracuse	EyeDisease

Resultados depois dos ajustes

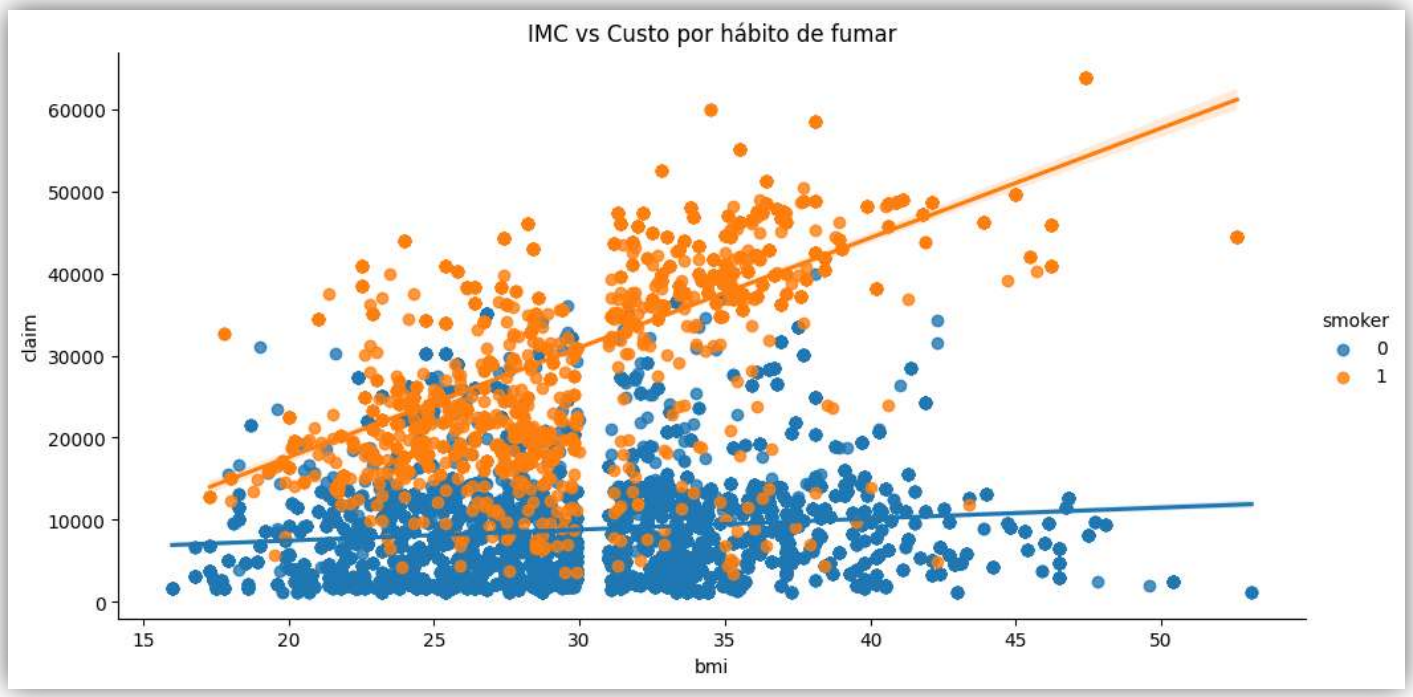
R^2: 0.9609072515534081

MAE: 482.1655767914533

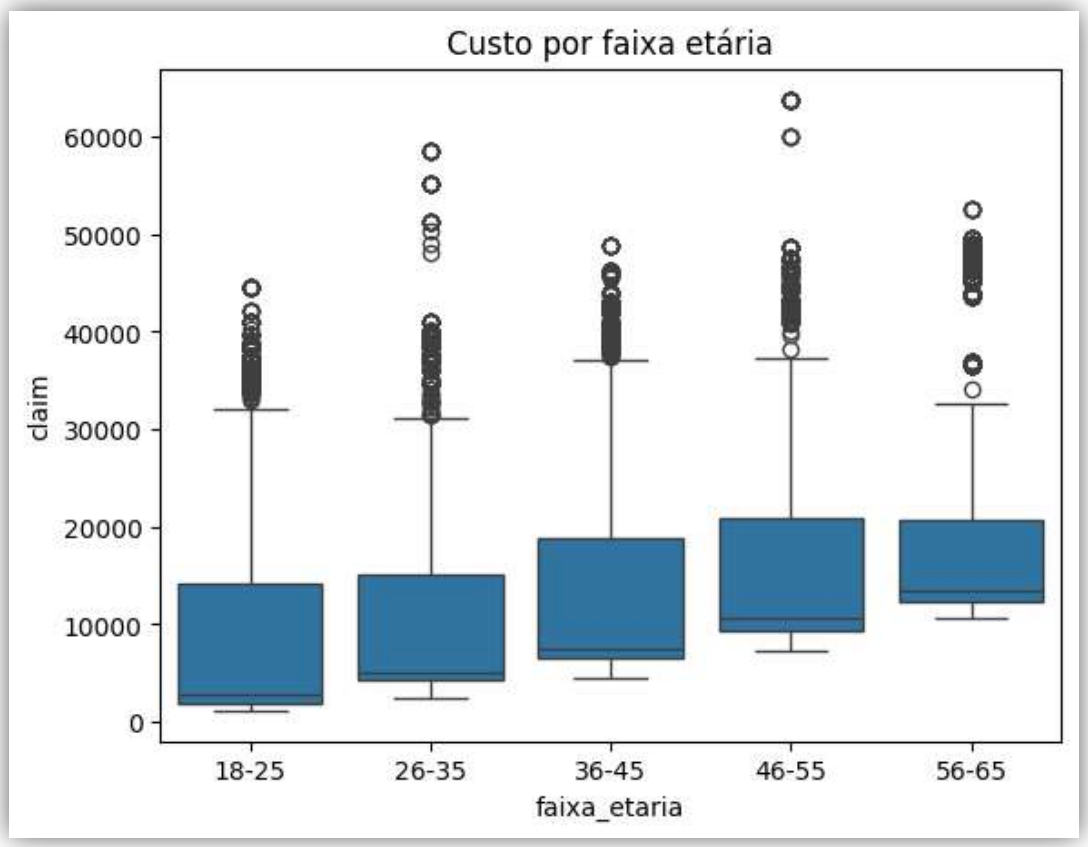
RMSE: 2377.8222727712287



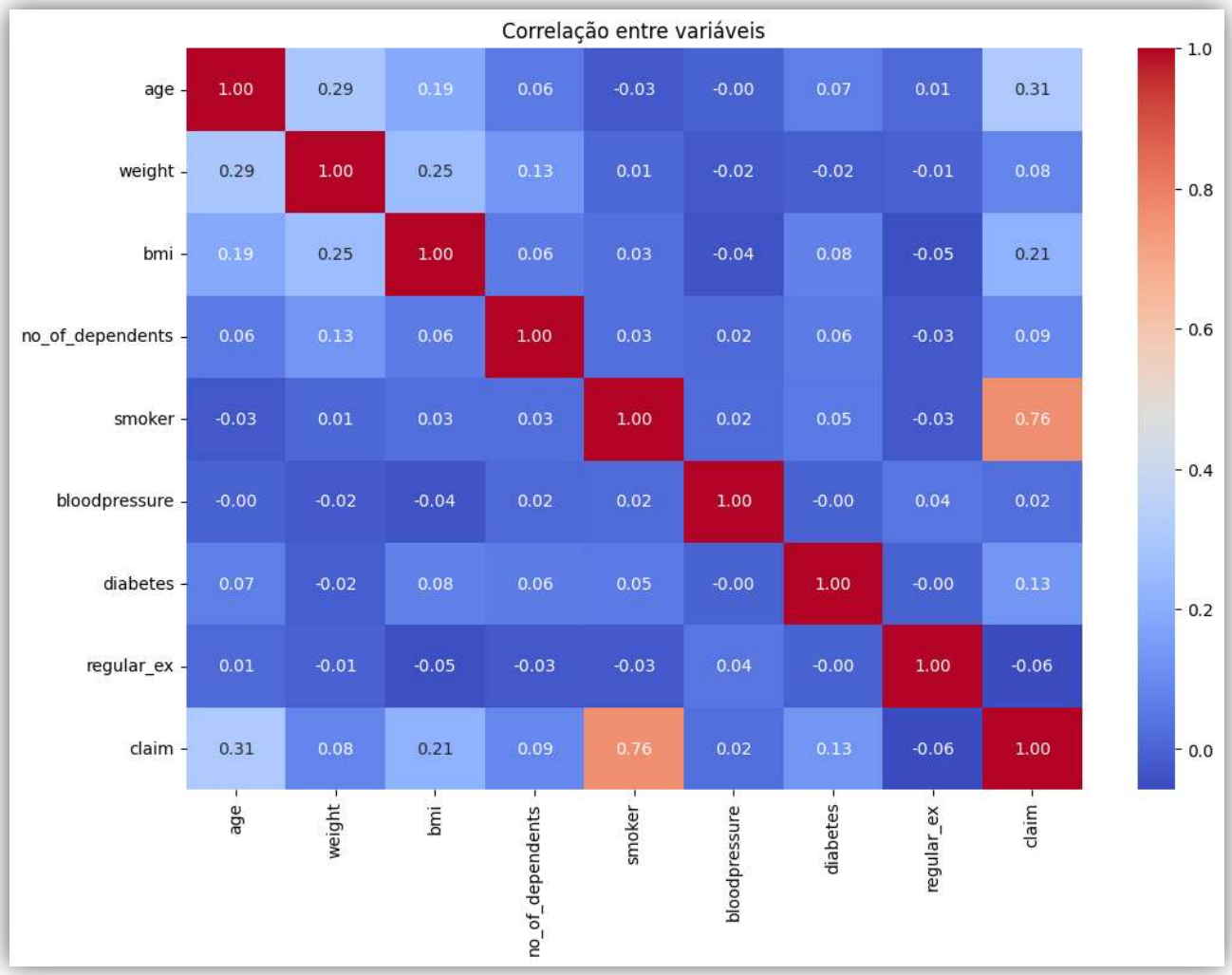
O gráfico de barras Custo médio por região (Top 20 cidades) apresenta uma comparação dos custos médios de seguro entre as 20 cidades com maior número de dados. As barras representam a média, enquanto as linhas verticais indicam a variabilidade dos custos em cada localidade.



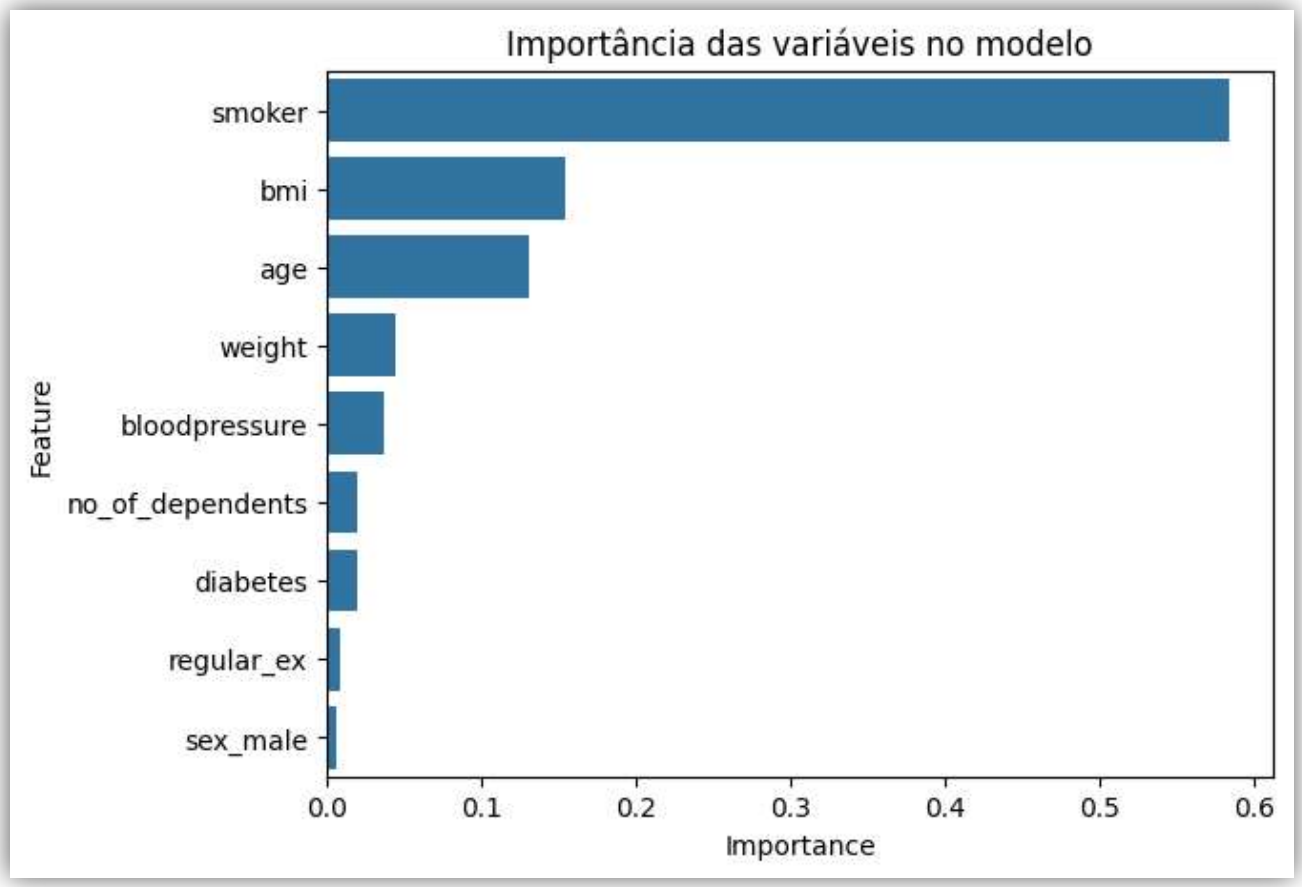
O gráfico IMC vs Custo por hábito de fumar analisa a relação entre o Índice de Massa Corporal (IMC) e os custos médicos, diferenciando entre fumantes e não fumantes. As linhas de tendência ajudam a visualizar como o IMC influencia os custos em cada grupo.



O boxplot Custo por faixa etária compara os custos médicos entre diferentes grupos de idade, permitindo a análise de variações nos gastos com saúde ao longo do ciclo de vida.



O mapa de calor Correlação entre variáveis apresenta a correlação entre todas as variáveis do conjunto de dados. As cores e os valores numéricos indicam a força e a direção das relações lineares entre as variáveis, auxiliando na identificação de padrões e dependências.



O gráfico de barras Importância das variáveis no modelo ilustra a relevância de cada variável na previsão dos custos médicos. As variáveis com barras mais longas têm maior impacto na precisão do modelo preditivo.