

**Base de dados:** Melbourne Housing Snapshot criada por Toni Pino.

Link: <https://www.kaggle.com/dansbecker/melbourne-housing-snapshot>

**Objetivo:** Realizar a predição dos preços das casas do DataSet Melbourne Housing Snapshot baseado nos dados fornecidos.

**Roteiro:**

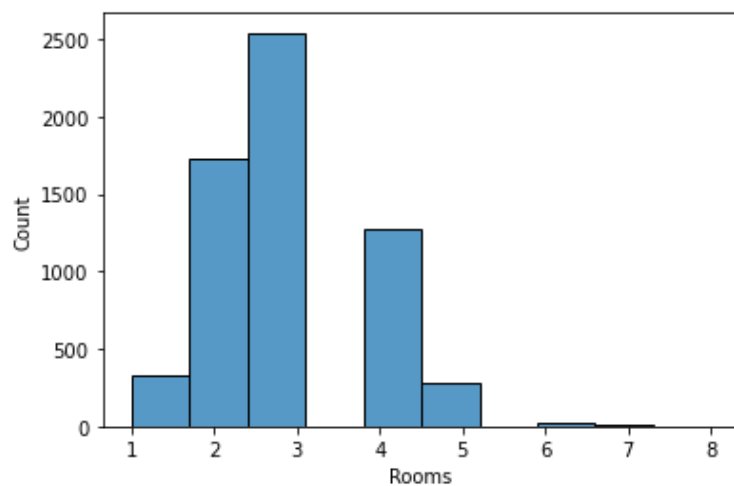
- Iniciaremos com uma análise dos campos do DataSet;
- Seguiremos com uma análise exploratória do cruzamento dos campos do DataSet;
- Finalizaremos com a predição do valor da casa baseado nos campos do DataSet.

### **Análise dos Campos do DataSet**

- **Suburb:** Nome do subúrbio.
- **Address:** Endereço da propriedade.
- **Rooms:** Cômodos.
- **Type:** Tipo da casa.
- **Price:** Preço da propriedade.
- **Method:** Método de venda.
- **SellerG:** Nome do vendedor.
- **Date:** Data da venda.
- **Distance:** Distância da propriedade em relação ao centro.
- **Postcode:** Código Postal.
- **Bedroom2:** Número de quartos (?).
- **Bathroom:** Número de banheiros.
- **Car:** Número de vagas para carro.
- **Landsize:** Área da propriedade.
- **BuildingArea:** Área construída.
- **YearBuilt:** Ano de construção.
- **CouncilArea:** Governo da área.
- **Lattitude:** Latitude.
- **Longitude:** Longitude.
- **Regioname:** Nome da região.
- **Propertycount:** Número de propriedades no subúrbio.

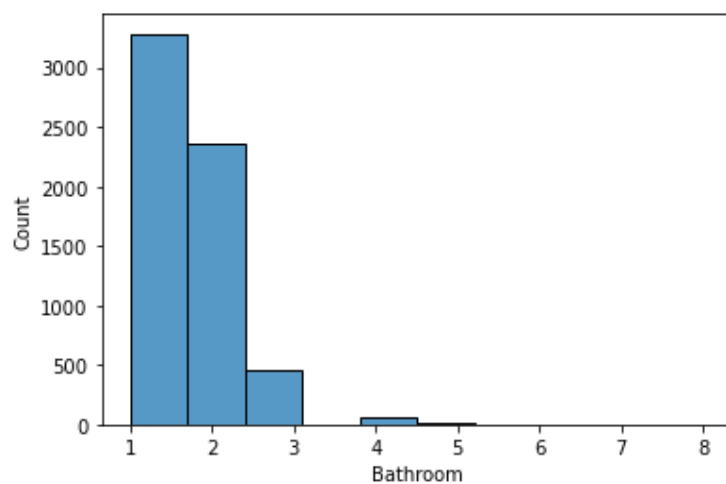
São 21 campos no total sendo 13 numéricos e 8 campos de texto. O DataSet possui 13580 entradas.

Vamos passar agora para a apresentação de algumas estatísticas dos campos listados acima.



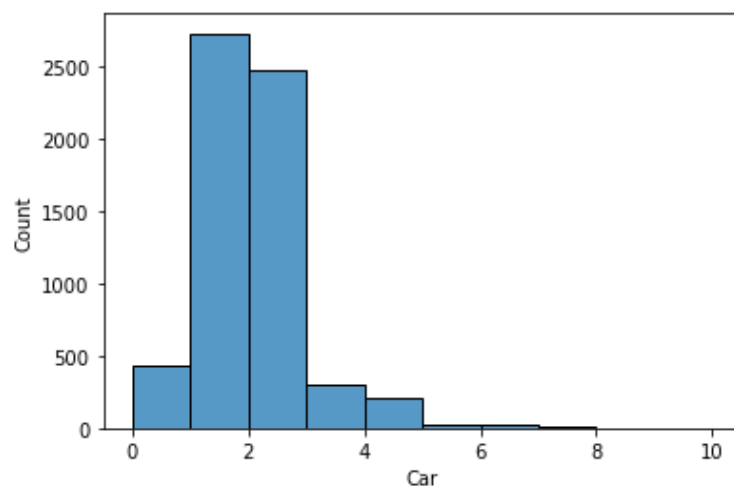
**Figura 1: Distribuição do número de quartos.**

A média do número de quartos é de 2.9 quartos por casa.



**Figura 2: Distribuição do número de banheiros.**

A média do número de banheiros é de 1.5 banheiros por casa.



**Figura 3: Distribuição do número de vagas para carro.**

A média do número de vagas é de 1.6 vagas de carro por casa.

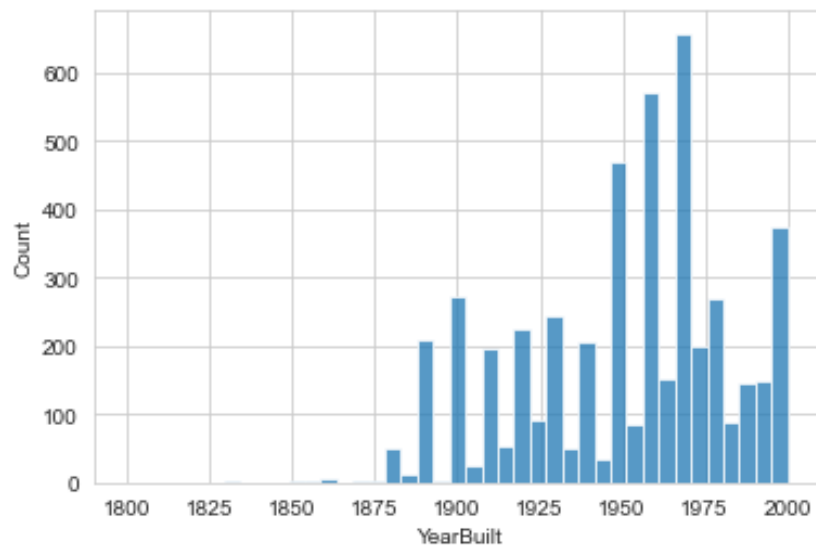


Figura 4: Distribuição do ano de construção das casas.

A cidade de Melbourne foi **fundada em 1835**, e desde esse ano o gráfico apresenta uma tendência de alta. Interessante notar que existem anos com um crescimento elevado no número de casas construídas e em sequência a escassez de casas sendo construídas. Além disso existiu um crescimento vertiginoso no número de casas construídas nos pós Segunda Guerra Mundial. Nos pós Segunda Guerra Mundial houve um *boom* econômico conhecido como Era de ouro do capitalismo. Este evento ocorreu entre os anos de 1945 e 1970, talvez exista uma relação entre o número de casas construídas e este evento histórico. Na década de 1970 houve um período de recessão, nota-se uma grande diminuição no número de casas construídas nesta década, este evento pode ser um fator para o decaimento do número de casas construídas nesses anos.

Nota-se que a economia está possivelmente ligada ao ramo de **Real Estate Investment Trust** com a tendência de anos com a economia mais forte o número de casas construídas cresce, o que contrasta com os eventos ocorridos durante a epidemia de COVID-19 (2019 - Atual), ocorreu um crescimento grande no setor de construção civil, o que na minha opinião está ligado as taxas de juros baixas (SELIC atual de 2.75%). Não encontrei dados históricos da taxa de juros da Austrália que vão além da década de 1970 para fazermos uma comparação. Creio que exista este contraste pois os bancos centrais diminuíram artificialmente as taxas de juros para incentivar o investimento, buscando limitar os danos causados pela crise.

Um ponto possivelmente interessante que não será possível analisar é o crash econômico de 2008 que teve muito a ver com o mercado de Real Estate Investment Trust, pois a base de dados contém registros até o final da década de 1990.



Figura 5: Distribuição dos preços das casas

O preço médio das casas é de U\$1.075.684, aqui podemos tirar o *insight* que meu deus, como é caro morar nessa cidade. Na figura 6 está a distribuição boxplot.

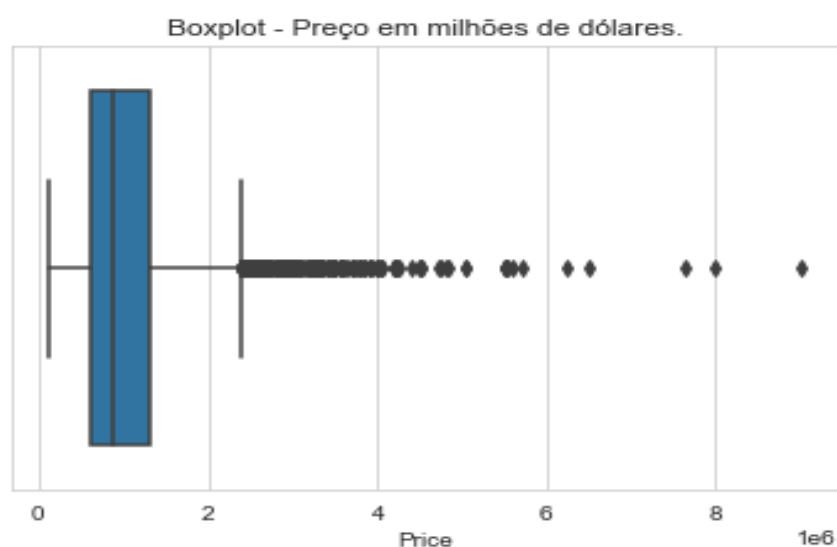


Figura 6: BoxPlot dos preços

Qual são os dados da casa com o maior valor de venda nesta base de dados?

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode
12094	Mulgrave	35 Bevis St	3	h	9000000.0	PI	Hall	29/07/2017	18.8	3170.0

Bathroom	Car	Landsize	BuildingArea	YearBuilt	CouncilArea	Lattitude	Longitude	Regionname	Propertycount
1.0	1.0	744.0	117.0	1960.0	Monash	-37.93168	145.16126	South-Eastern Metropolitan	7113.0

E por motivos de comparação, quais são os dados da casa com o menor valor de venda nesta base de dados?

Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	
1805	Caulfield	30 Pyne St	4	h	131000.0	PI	Rodney	25/02/2017	8.9	3162.0

Bathroom	Car	Landsize	BuildingArea	YearBuilt	CouncilArea	Lattitude	Longitude	Regionname	Propertycount
1.0	2.0	499.0	155.0	1920.0	Glen Eira	-37.8864	145.0242	Southern Metropolitan	2379.0

Esse preço está muito fora do preço plausível pela casa, seria alguém lavando dinheiro?

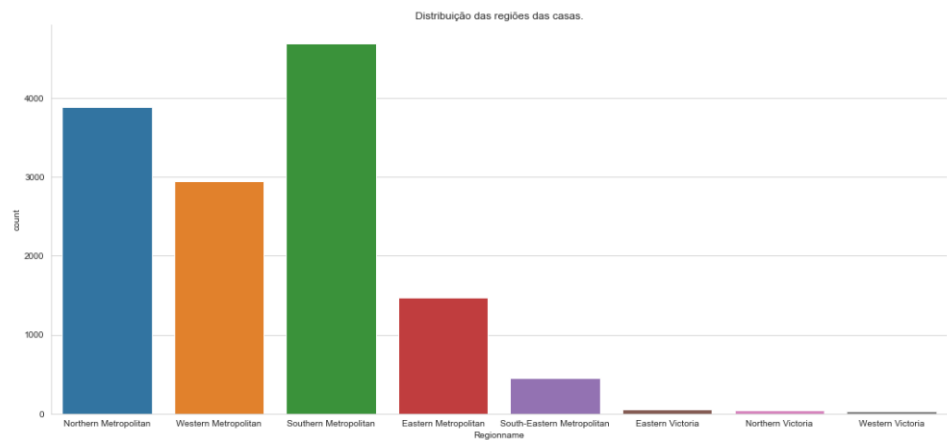


Figura 7: Distribuição das casas por região.

A figura 7 mostra a distribuição das casas por região, nota-se a concentração nas regiões de Northern Metropolitan, Western Metropolitan e Southern Metropolitan.

Análise Exploratória dos Campos

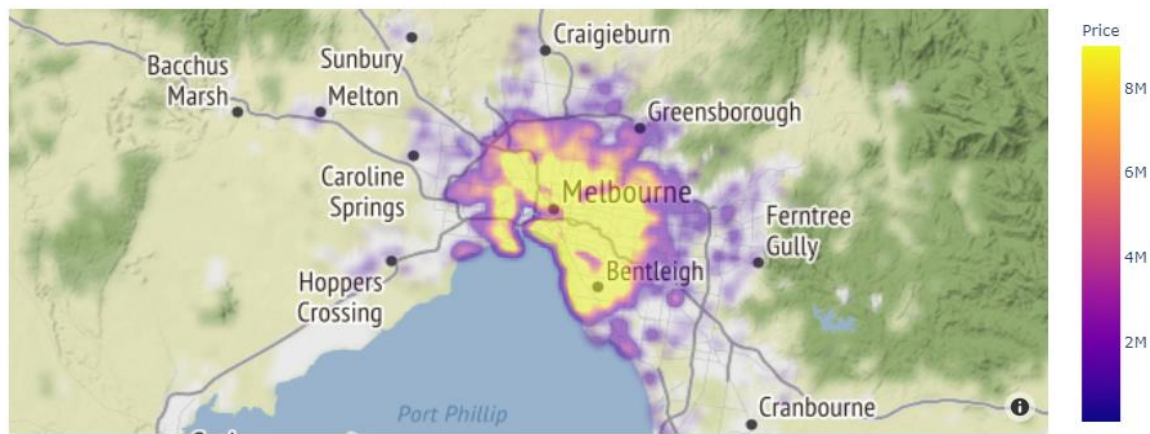


Figura 8: Heatmap Casa x Preço

A figura 8 mostra um mapa de calor da distribuição das casas em relação ao preço na cidade de Melbourne, Austrália. Por premissa existem certos campos que possivelmente adicionam valor na casa, como por exemplo os campos de **número de cômodos, quartos e banheiros**, além dos campos de **área total, área construída** e o **número de vagas para carros**. Vamos utilizar um gráfico de barras para ver se isso realmente ocorre.

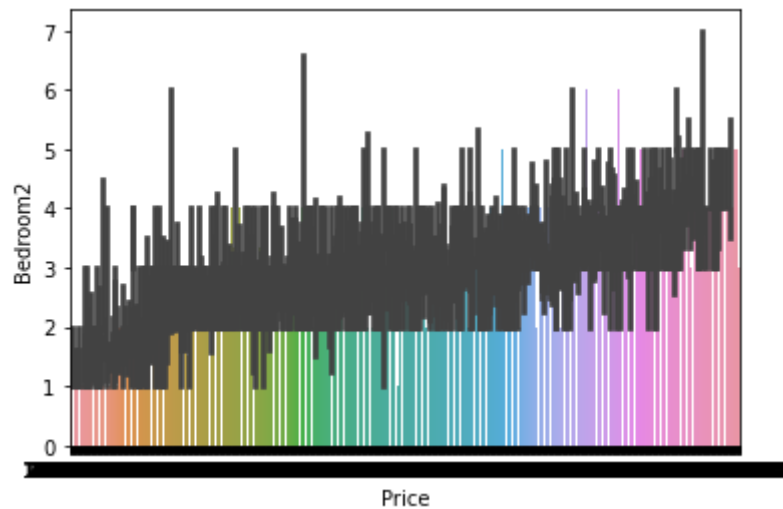


Figura 9: Preço x Número de quartos

Note que existe uma certa relação entre o preço com o número de quartos da casa.

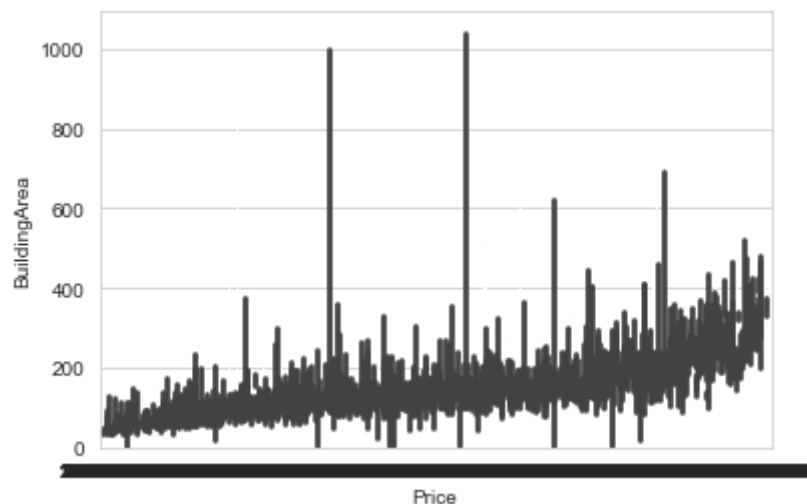
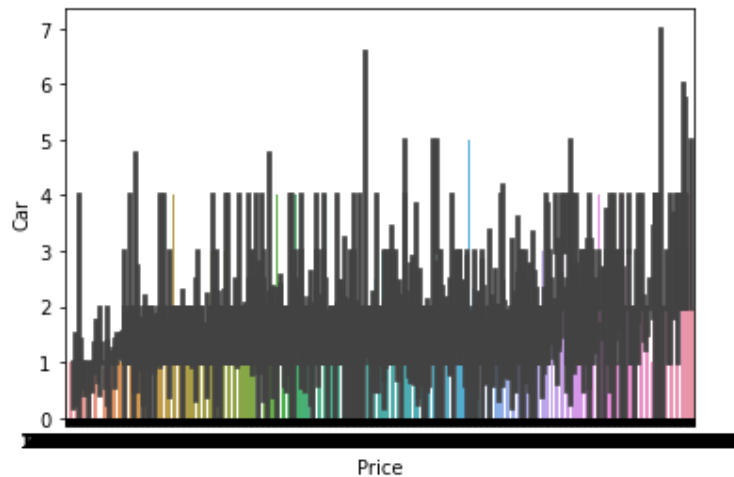


Figura 10: Preço x Área construída

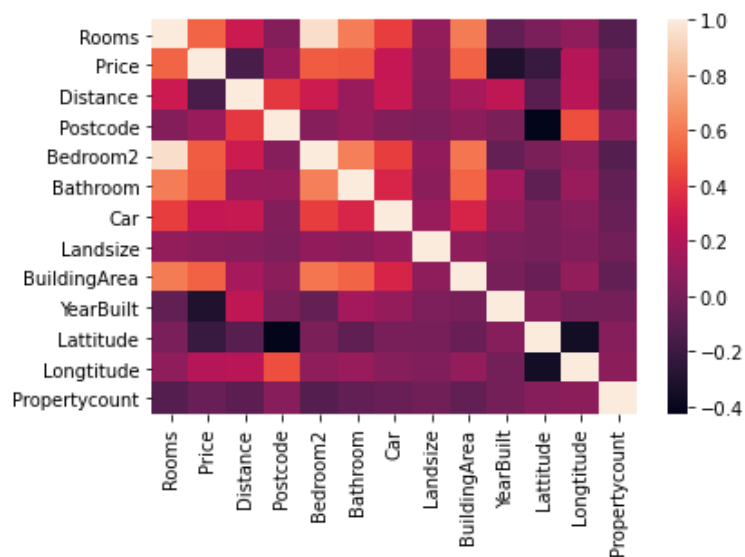
Também existe uma relação entre o preço com a área construída da casa.



**Figura 11: Preço x Vagas para carro**

Aparentemente não existe uma relação clara entre o preço da casa com o número de vagas para carro.

Para finalizar as análises de correlação, segue uma figura com o mapa de calor de correlação das variáveis.



**Figura 12: HeatMap de correlação**

### Predição dos preços

Será utilizado Regressão Linear para realizar a predição dos preços das casas de Melbourne. As variáveis utilizadas serão: Rooms, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, PropertyCount. Será utilizado o sklearn para prover o modelo de regressão e as ferramentas de tratamento de dados.

```

from sklearn.model_selection import train_test_split #importando o método utilizado para separar em modelo de treino e teste

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,random_state=101)

from sklearn.linear_model import LinearRegression #importando o modelo de Regressão Linear

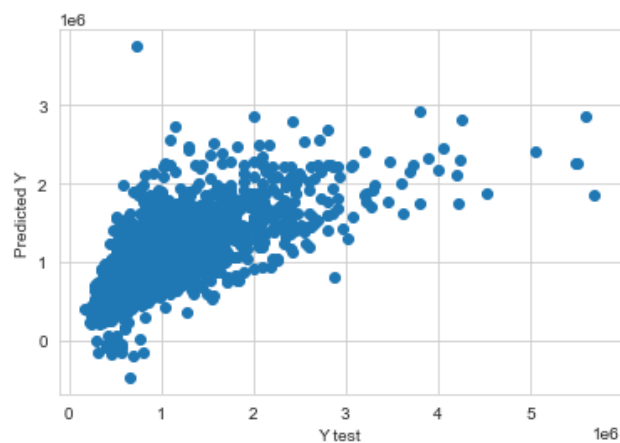
lm = LinearRegression() #instanciando o modelo de regressão linear

lm = LinearRegression() #instanciando o modelo de regressão linear

lm.fit(X_train, y_train)

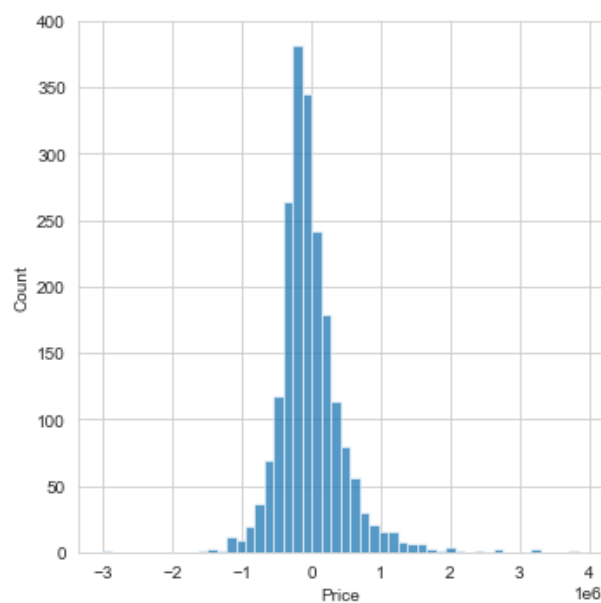
LinearRegression()

```



**Figura 13: Predito x Preço**

A figura 13 mostra a relação entre o preço que foi predito e o preço de fato da casa.



**Figura 14: Resíduos entre Preço real – Preço predito**



Note que existe uma concentração forte nos resíduos próximos de 0, o que significa que o preço está alinhado com o preço predito.

Coefs	
Rooms	212908.044470
Distance	-39489.905229
Bedroom2	17590.334761
Bathroom	163368.714397
Car	36900.543761
Landsize	19.413011
BuildingArea	2072.961850
Propertycount	-1.298125

Figura 15: Coeficientes de predição

### Como interpretar tais dados?

Para cada aumento de uma unidade, haverá um acréscimo do coeficiente no valor total predito da casa.

**Exemplo:** Para cada quarto é adicionado no preço final da casa U\$212.908, para cada quilômetro a mais de distância em relação ao centro é subtraído cerca de U\$39.489 do preço final da casa, e assim por diante.

Vamos agora predizer o preço de uma casa aleatória do conjunto de dados, aleatoriamente escolhemos o valor 25.

Número de quartos: 2, Distância do centro: 2.5, Bedroom2: 2, Bathroom: 1, Car: 0, Landsize: 147, BuildingArea: 85, PropertyCount: 4019.

Exemplo do cálculo do preço da casa de número 25:

**Rooms 3:**  $3 * 212.908 = 638.724$  dólares

**Distance 3.3:**  $3.3 * 39.489 = -130.313$  dólares

**Bedroom2 3:**  $3 * 17.590 = 52.770$  dólares

**Bathroom 2:**  $2 * 163.368 = 326.736$  dólares

**Car 1:**  $1 * 36.900 = 36.900$  dólares

**Landsize 211:**  $211 * 19.41 = 4095$  dólares

**BuildingArea 147:**  $147 * 2073 = 304.731$  dólares

**PropertyCount:**  $3280 * -1.3 = -4264$  dólares

**Preço estimado:** 1.229.379 Dólares

**Preço estimado X Preço real**

Preço estimado: 1.229.379 dólares

Preço real: 1.315.000