

2º Projeto Prático – Dados

Análise de Dados e Predição com Python e Pandas

João Daniel Temporin

Relatório

Primeiramente, ao importar o dataset desse projeto, consultei o tamanho da base de dados:

- **11.657 linhas;**
- **8 colunas** (Endereço, Bairro, Área (km²), Quartos, Garagem, Tipo, Aluguel (R\$) e Total)

Vamos utilizar o Aluguel (R\$) como variável dependente (Y). Observando o Boxplot abaixo (Figure 1), vemos que nossa mediana se encontra na faixa dos 2500. Para chegar no valor exato utilizei o `.median()` da biblioteca Pandas, resultando em uma mediana de 2.415. Também podemos observar que há bastante outliers, incluindo um valor extremamente discrepante de 25.000. Para que nossa Regressão Linear fique mais ajustada, realizei a exclusão desse valor.



Figure 1. Boxplot da variável Aluguel.

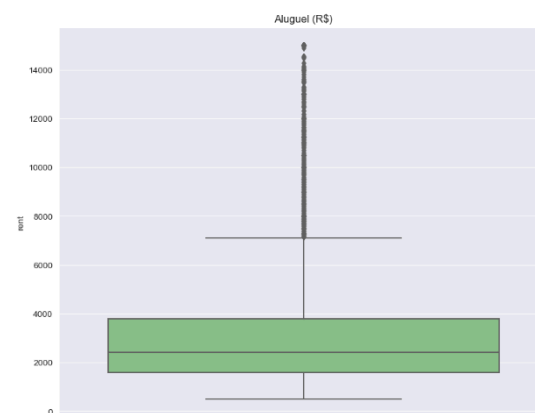


Figure 2. Boxplot após remoção do Outlier extremo.

Plotei outros dois Boxplots com a variável dependente Aluguel, mas desta vez utilizando as duas variáveis independentes Garagem e Quartos. Notamos uma tendência de aumento no valor do aluguel conforme é maior o número de garagens (Figure 3). Os valores também tendem a se tornar menos discrepantes, observando-se através da diminuição das “caudas”. As mesmas análises se aplicam a variável Quartos, com exceção as diferenças entre os boxplots de 5 e 6 quartos, que são quase idênticas. (Figure 4).

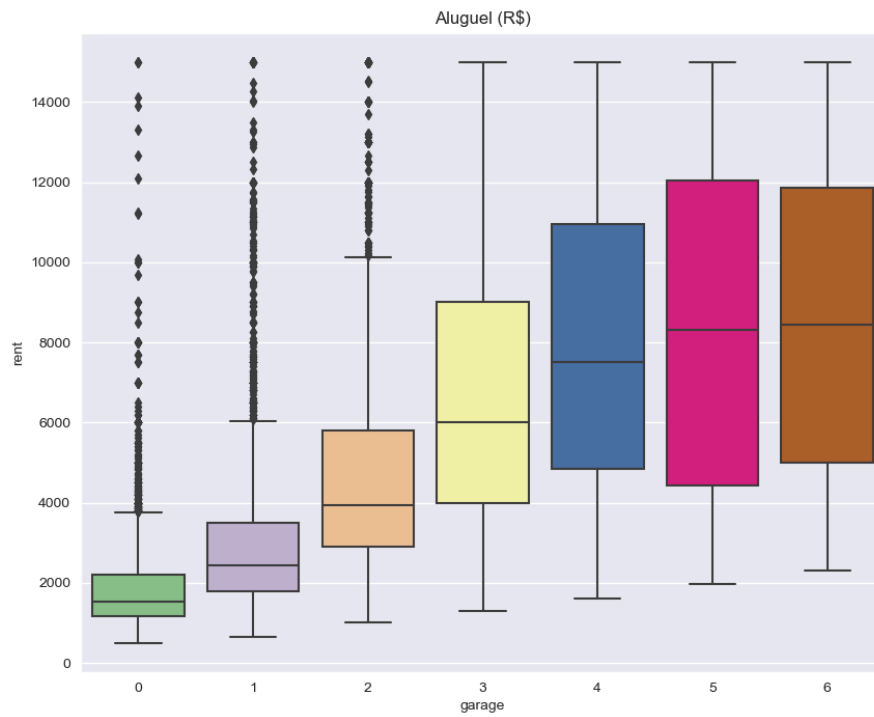


Figure 3. Boxplot da variável Aluguel em conjunto com a variável Garagem.

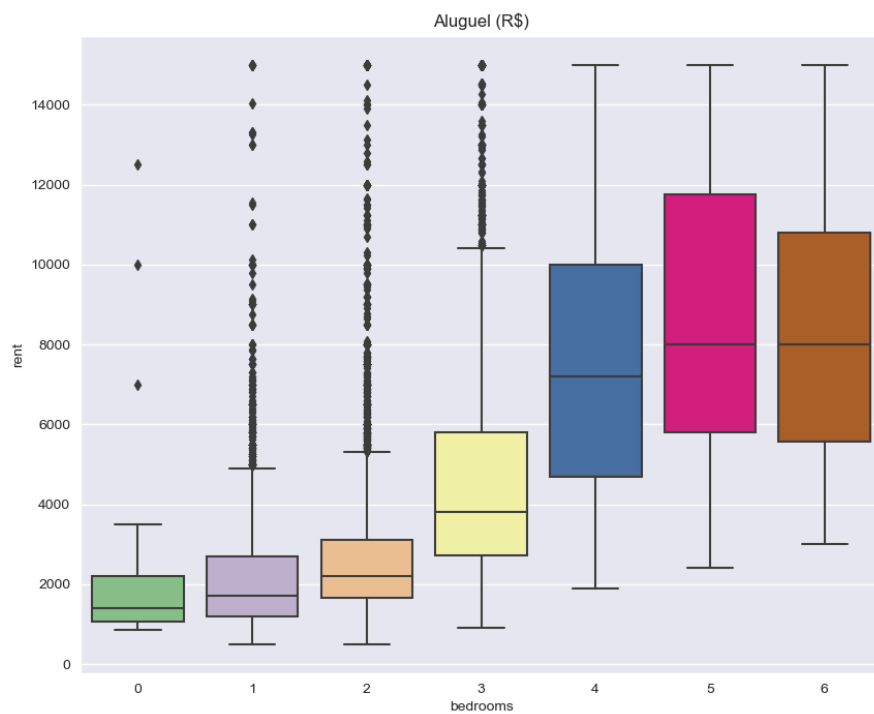


Figure 4. Boxplot da variável Aluguel em conjunto com a variável Quartos.

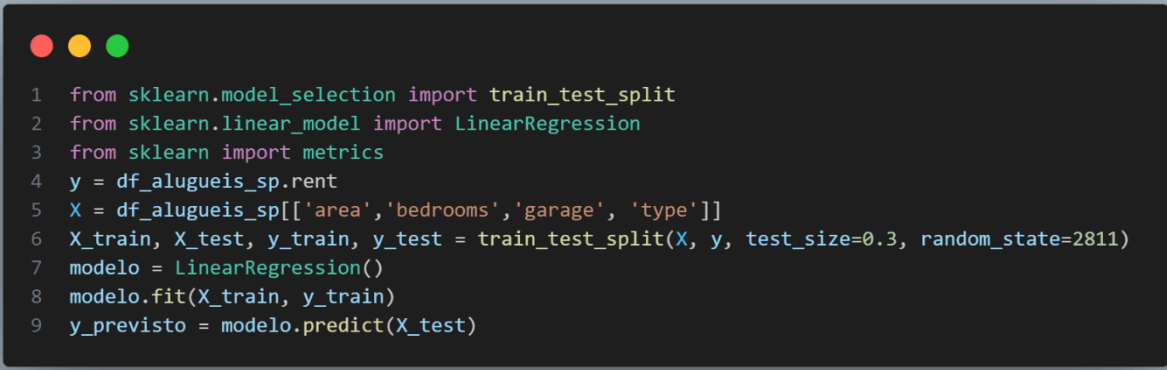
Para ajustar mais os dados para calcular a Regressão Linear, apliquei o método de Label Encoding na variável Tipo, sendo:

- Apartamento = 0
- Casa = 1
- Casa em condomínio = 2
- Studio e kitnet = 3

Com isso, será possível utilizar uma variável que era do tipo categórica, como numérica. O tipo do imóvel é de suma importância em nosso modelo de Regressão pois há uma variação entre essas categorias em relação ao valor do aluguel (por exemplo, o aluguel de um apartamento tende a ser maior de uma casa com a mesma quantidade de garagens, quartos e área).

Usando a biblioteca sklearn para calcular a Regressão Linear, utilizei as variáveis Área, Quartos, Garagem e Tipo como X (Figure 5).

Calculando a Correlação de Pearson, chegamos a um valor de 0.49 (0.48 para os dados de teste). Isso significa que temos uma relação moderada entre as variáveis e que conforme as variáveis independentes aumentam, a variável dependente também aumenta (conforme observado nos Boxplots).



```

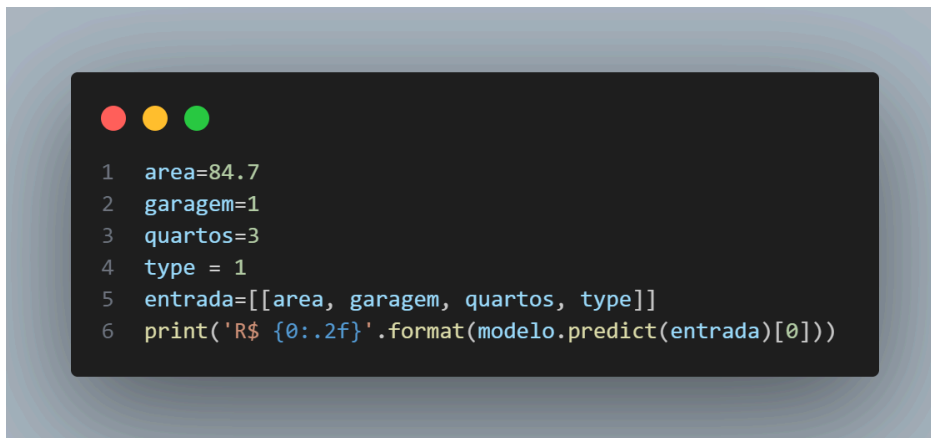
1 from sklearn.model_selection import train_test_split
2 from sklearn.linear_model import LinearRegression
3 from sklearn import metrics
4 y = df_alugueis_sp.rent
5 X = df_alugueis_sp[['area', 'bedrooms', 'garage', 'type']]
6 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=2811)
7 modelo = LinearRegression()
8 modelo.fit(X_train, y_train)
9 y_previsto = modelo.predict(X_test)

```

Figure 5. Regressão Linear utilizando a biblioteca SKLearn.

Para utilizar os dados previstos de forma analítica, utilizamos um simulador no qual retorna o valor previsto da variável Y a partir dos números que inserirmos nas variáveis X (Figure 6). Nesse exemplo, em uma casa com 84.7 de área, 1 garagem, 2 quartos e tipo 1 (Casa), foi previsto um aluguel de R\$ 4.339,77.





```

1 area=84.7
2 garagem=1
3 quartos=3
4 type = 1
5 entrada=[[area, garagem, quartos, type]]
6 print('R$ {:.2f}'.format(modelo.predict(entrada)[0]))

```

Figure 6. Simulador.

Principais análises:

- O aumento no número de garagens impacta de forma mais minuciosa no valor do Aluguel (média de R\$ 13 de aumento);
- A quantidade de quartos impacta bastante no aumento do Aluguel (cerca de R\$ 500), sendo uma variável importante a se analisar quando se busca investir em uma propriedade em São Paulo.
- Apartamentos geralmente possuem um aluguel mais elevado quando comparado a casas;