2º Projeto Prático – Dados

Análise de Dados e Predição com Python e Pandas

Análises e conclusões

Após avaliar os resultados obtidos do conjunto de dados de aluguel de São Paulo, podemos ter as seguintes conclusões:

1. O conjunto de dados não apresenta valores nulos e não apresenta dados duplicados.
2. Foram removidos alguns outliers:
    - Registros com área igual a zero;
    - Registro com valor de aluguel de R$ 25.000,00, por se tratar de uma kitnet de 24m²;
    - Registros com número de quartos igual a zero, uma vez que a maior parte dos imóveis do tipo "Studio e kitnet" apresenta 1 quarto.
    - Não foram removidos outliers a partir da análise do boxplot, uma vez que se tratavam de valores de área e de valor do aluguel considerados condizentes com a realidade.
3. Foram obtidos os valores de média, mediana e desvio padrão para as colunas de Área, Valor do aluguel e Valor total:

	Média	**Mediana**	**Desvio padrão**
Área	84,40	60,00	74,06
Valor do aluguel	3252,32	2426,00	2643,35
Valor total	4082,92	3060,00	3347,80

4. De acordo com os valores de correlação, podemos analisar que todas as associações são positivas, conforme esperado. Já que sabemos que a área de um imóvel tende a aumentar com o aumento do número de quartos e o número de vagas de garagem. Além disso, o valor do aluguel e o valor total das despesas também tendem a ser maiores de acordo com a área do imóvel. Para todas as relações entre as variáveis, podemos observar uma relação maior de 50%. O menor valor das relações observado é de 0,5291 (relação entre número de quartos e valor total) e o maior valor das relações é de 0,9781, que corresponde à relação entre aluguel e total.

5. Os gráficos de distribuição de frequências de área e do valor do aluguel apresentaram assimetria à direita, o que indica que existem alguns registros com valores de área e aluguel que se distanciam da média. Ou seja, poucos imóveis tendem a ser grandes e com alto custo de aluguel, enquanto a maioria se agrupa em torno de 84m² e R$ 2.000,00, locais onde ocorrem os picos de densidade dos gráficos de área e valor do aluguel, respectivamente.

6. A mesma assimetria à esquerda também é observada para o gráfico de distribuição de frequências do valor total, conforme o esperado.

7. Nos gráficos de dispersão entre as variáveis aluguel e total (eixo y) em função da área (eixo x), vemos um comportamento muito disperso, o que indica que não há correlação forte entre as variáveis.
8. Nos gráficos de dispersão entre as variáveis aluguel e total (eixo y) em função do número de quartos e número de garagens (eixo x), vemos as dispersões organizadas em quantidades que vão de 1 a 6 para os quartos e de 0 a 6 para as garagens. No gráfico total x quartos, vemos uma certa tendência de aumento do total até a quantidade de 4 quartos e que diminui a partir de 5 quartos. No gráfico total x garagens há uma tendência de aumento do total em função do aumento do número de vagas de garagem.

9. Foi traçada uma reta de regressão linear para o valor do aluguel x área a partir de um modelo de predição. Tanto a comparação visual entre a dispersão dos dados e a reta quanto o valor do R^2 (0,448) indicam um comportamento entre as variáveis que não pode ser representado satisfatoriamente pelo modelo.

10. Também foi desenvolvido um modelo de predição dos dados considerado os valores de aluguel e total em função da área, da quantidade de quartos e da quantidade de garagens. O valor obtido foi de 0,48. Dessa forma, o modelo apresentado é muito distante da realidade e não representa de forma razoável a relação dos valores de aluguel e total com a área, o número de quartos e o número de garagens. Para ser considerado um modelo que representa satisfatoriamente o conjunto de dados, o valor de R^2 deveria ser pelo menos de 0,95.

Considerações finais: Infelizmente, os modelos testados não representaram de forma significativa os dados reais. Dessa forma, não podemos afirmar que estes modelos poderiam ser aplicados para predição dos valores de aluguel em São Paulo. Uma sugestão é dividir o conjunto de dados por bairros e desenvolver novos modelos preditivos nos imóveis de cada bairro, uma vez que este é uma variável qualitativa que costuma ser de grande influência na precificação dos aluguéis de imóveis.