Grupo ZAP: Data Science Challenge

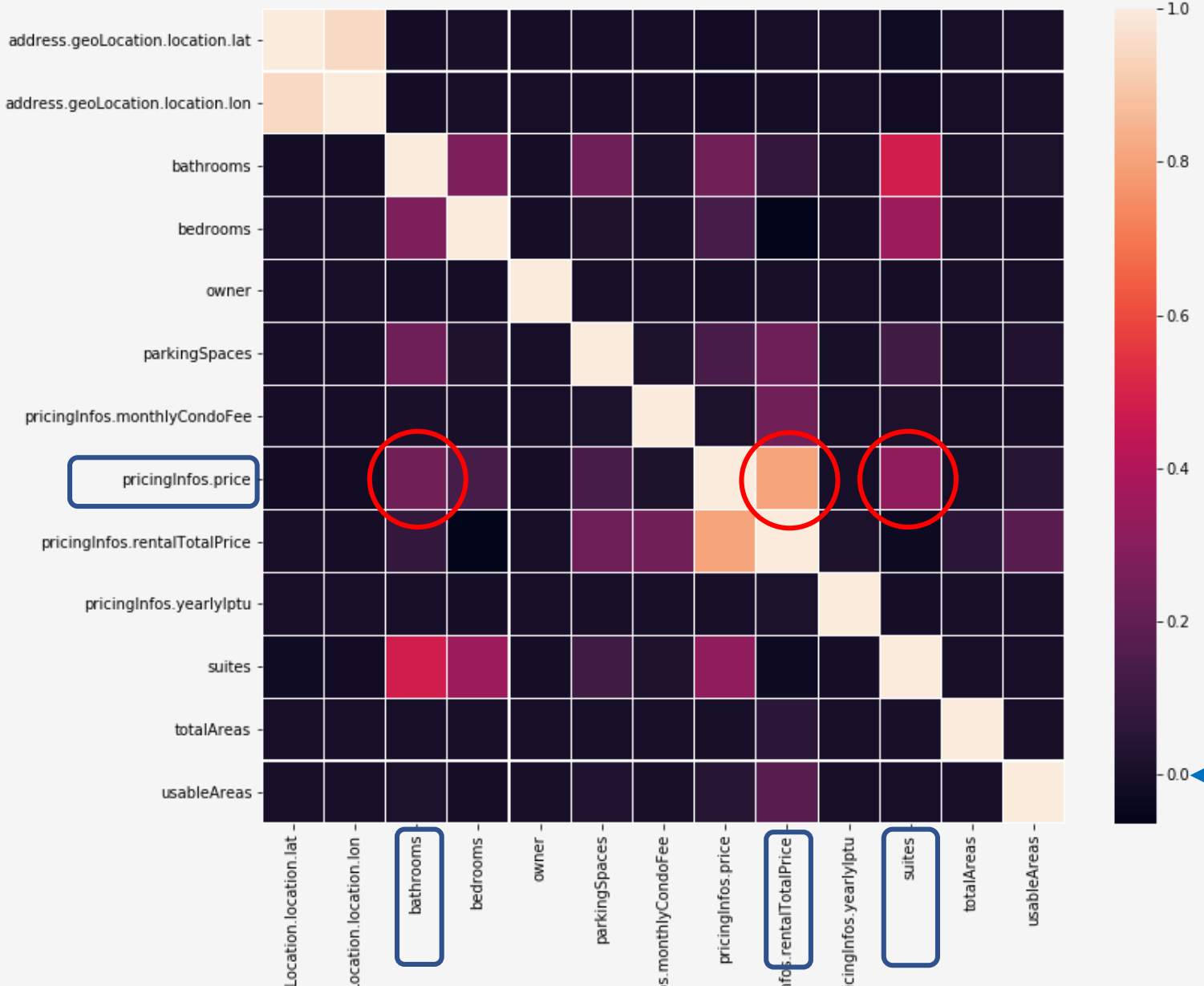
Candidato: Bruno B. Nasser

Desafio - Quanto custa?

Desenvolver uma maneira automática de estimar um preço de **venda** para os apartamentos no dataset de teste.

ANÁLISE DE DADOS E ELIMINAÇÃO DE OUTLIERS

MAPA DE CALOR PARA DADOS BRUTOS



Maior correlação

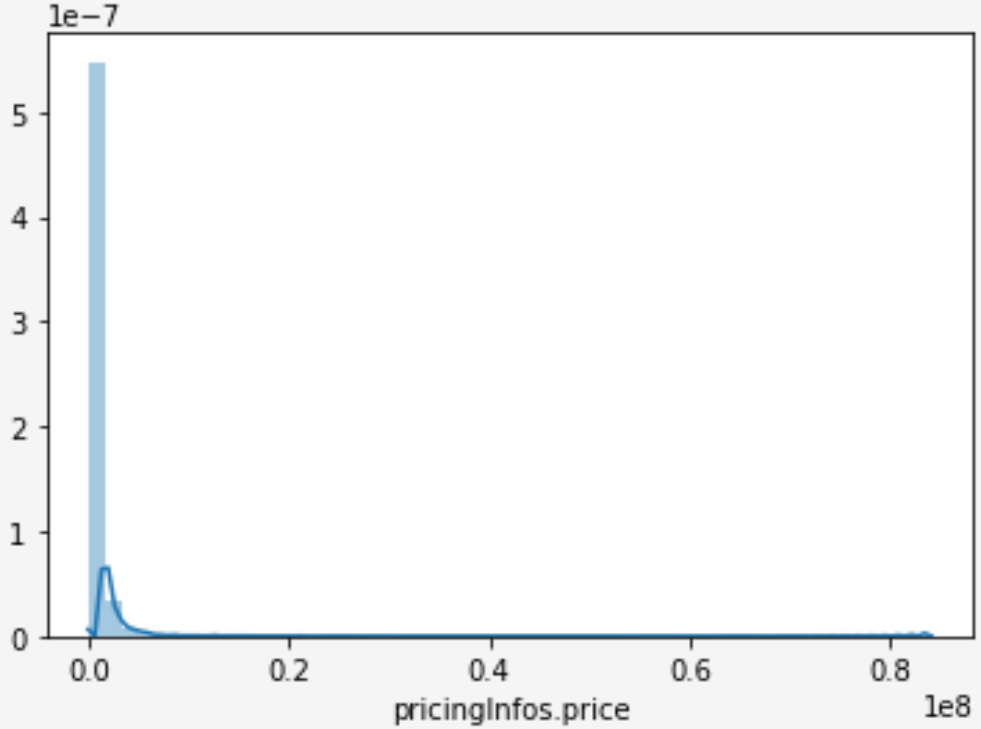
Menor correlação

Quadrados mais claros:

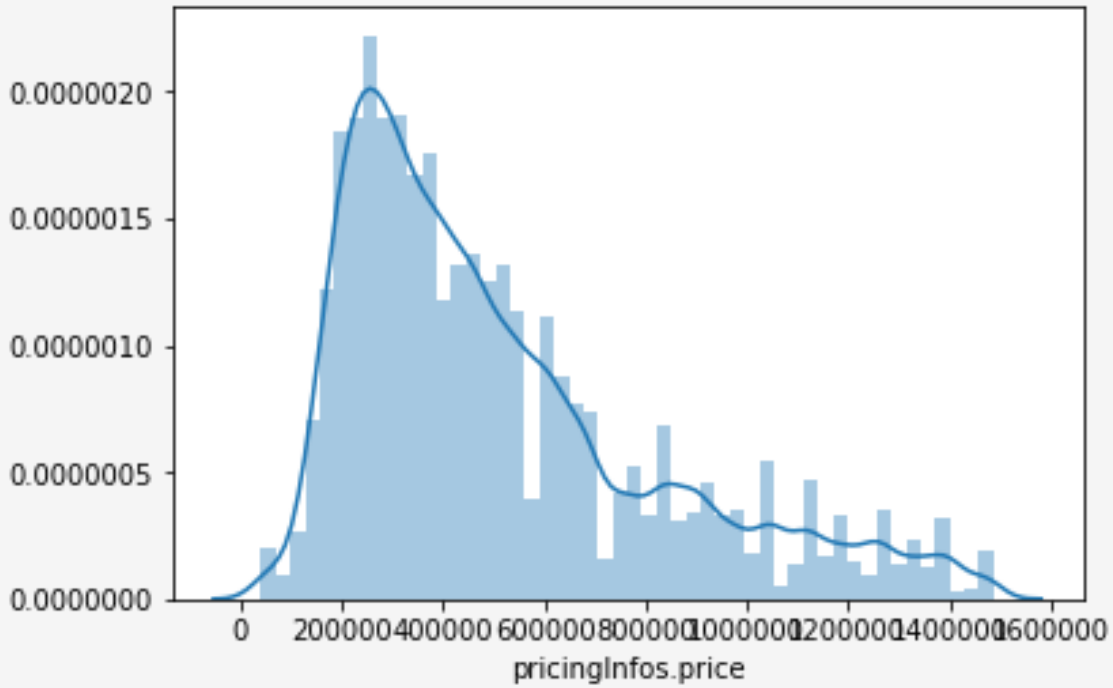
"Price" tem um fator de correlação maior com o valor dos alugueis, numero de suítes e banheiros. Em seguida vem numero de quartos e vagas na garagem.

Análise da variável “Price”

Histograma dos dados bruto



Histograma dos dados sem outliers



Análise: Os dados foram filtrados para 40.000<Price<1.487.500 de reais.

“Price” em relação ao número de banheiros

DADOS BRUTO

Gráfico de dispersão: Price x Bathrooms

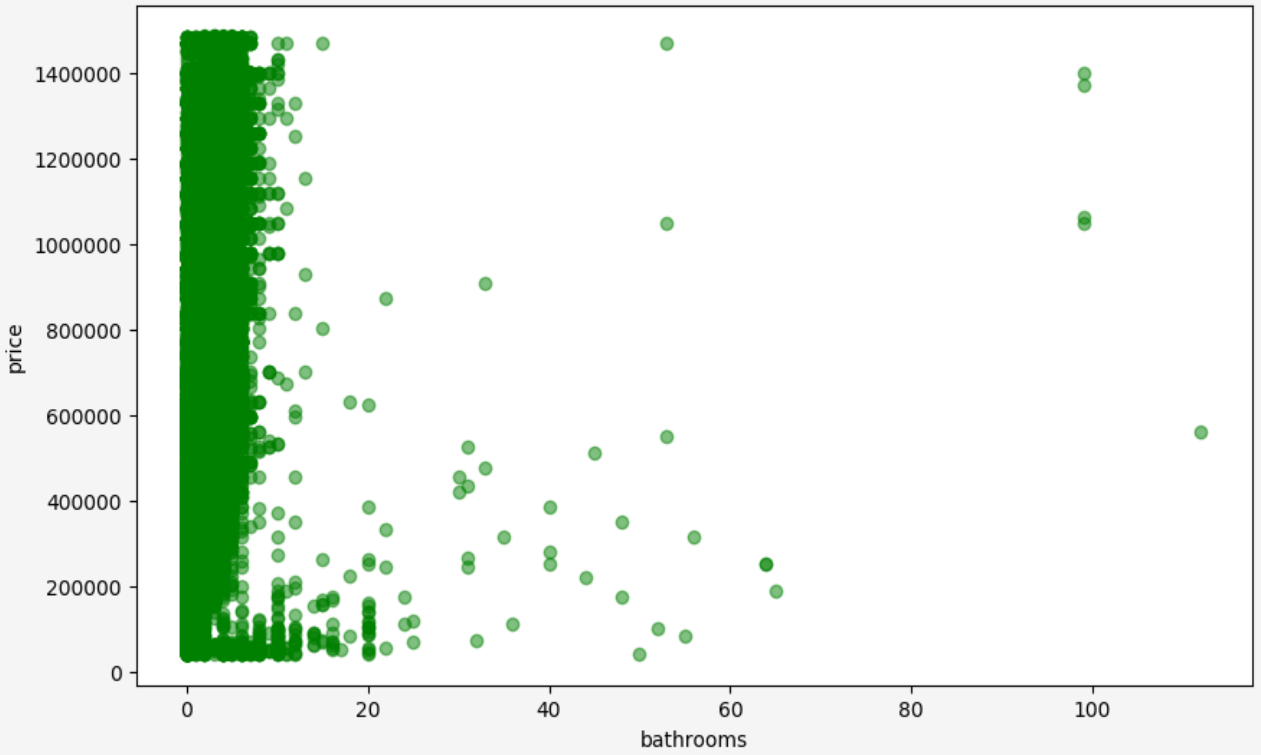
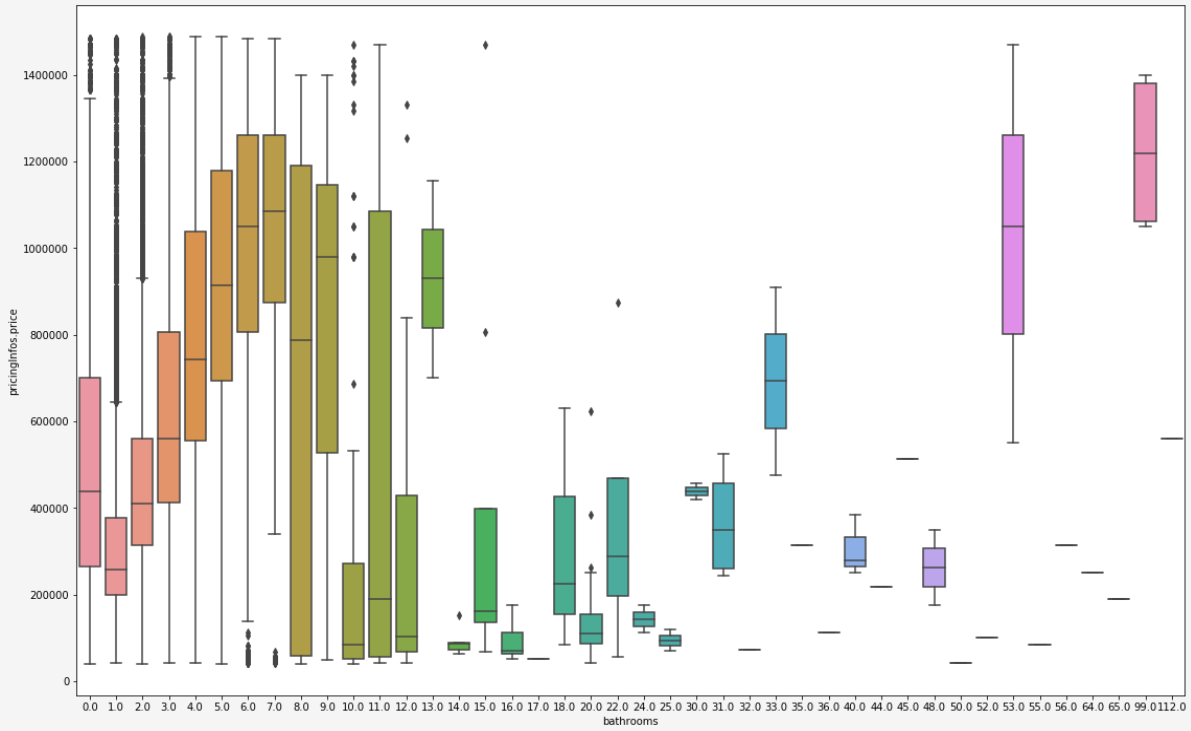


Diagrama de caixa: Price x Bathrooms



Análise: No diagrama de caixa a média de preço dos imóveis cresce à medida que aumenta o número de banheiros para até 7 banheiros, a partir disso a relação se distorce. Assim sendo optou-se por filtrar os dados entre a quantidade de 1 a 7 banheiros

“Price” em relação ao número de banheiros

DADOS SEM OUTLIERS

Gráfico de dispersão: Price x Bathrooms

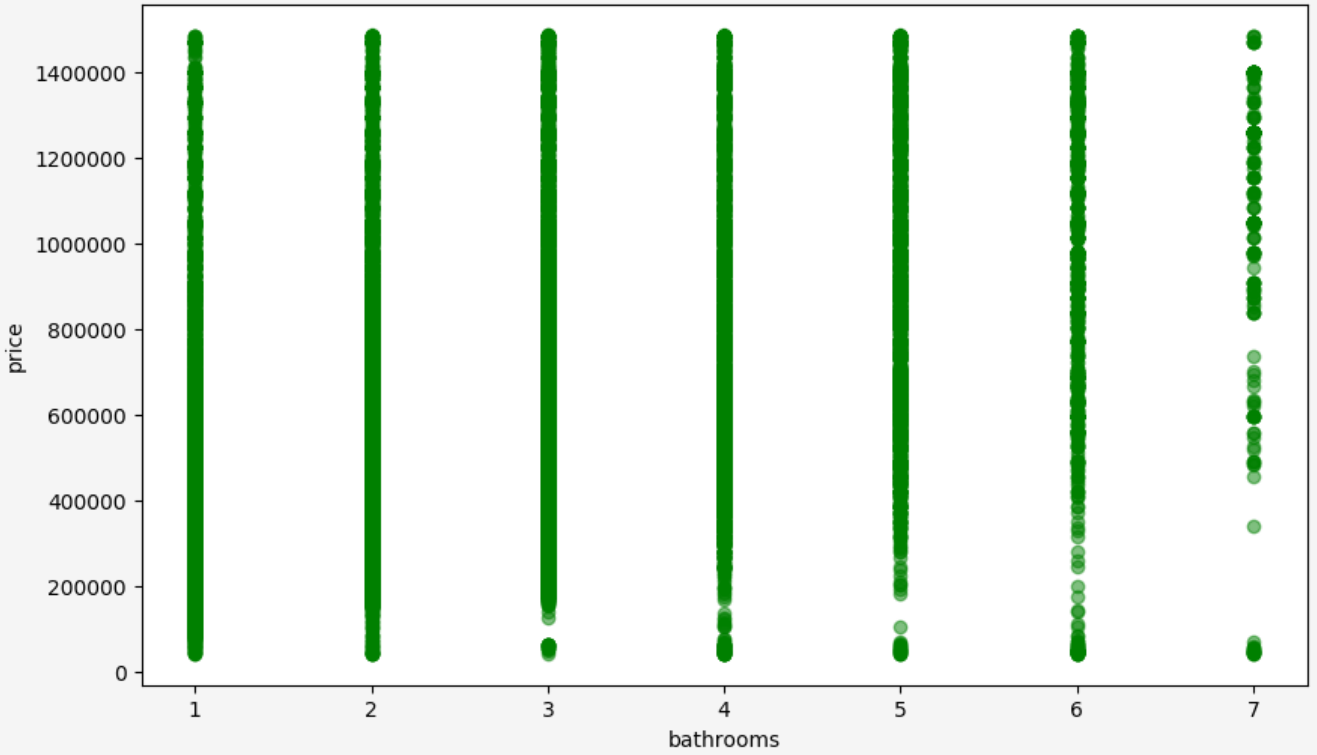
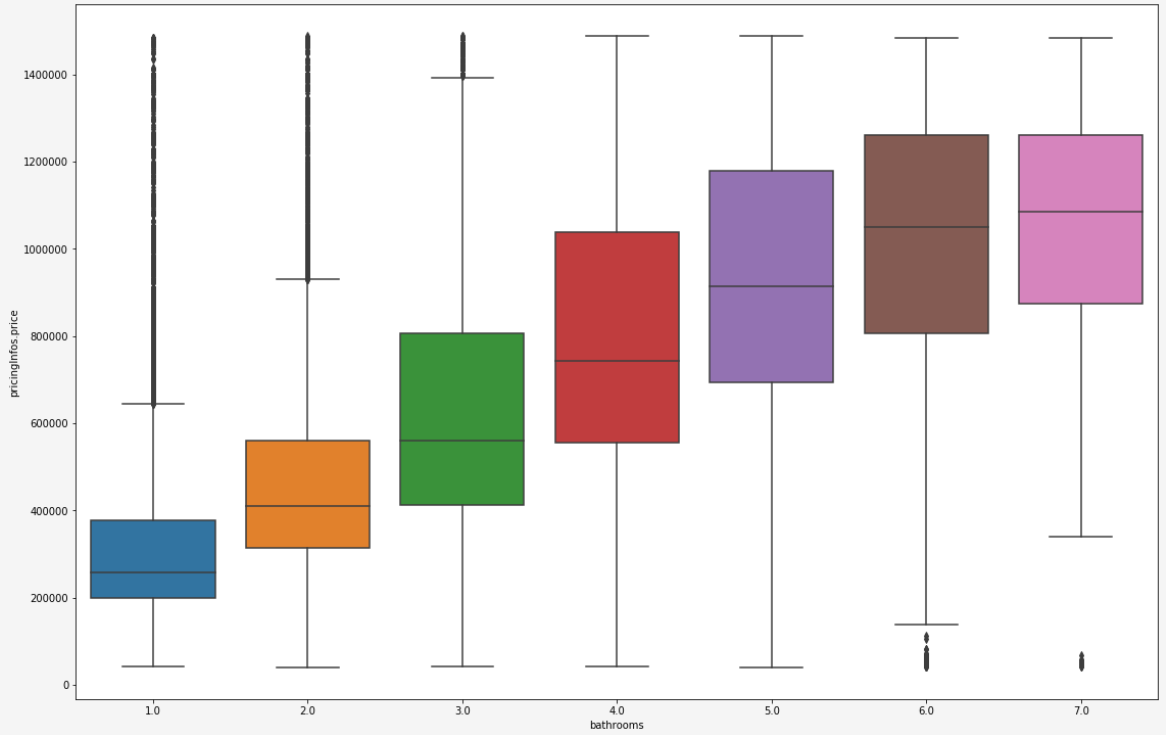


Diagrama de caixa: Price x Bathrooms



Análise: Sem outliers é possível observar uma melhor relação entre as variáveis.

“Price” em relação a suítes, quartos, vagas

DADOS SEM OUTLIERS

Diagrama de caixa: Price x Bedrooms

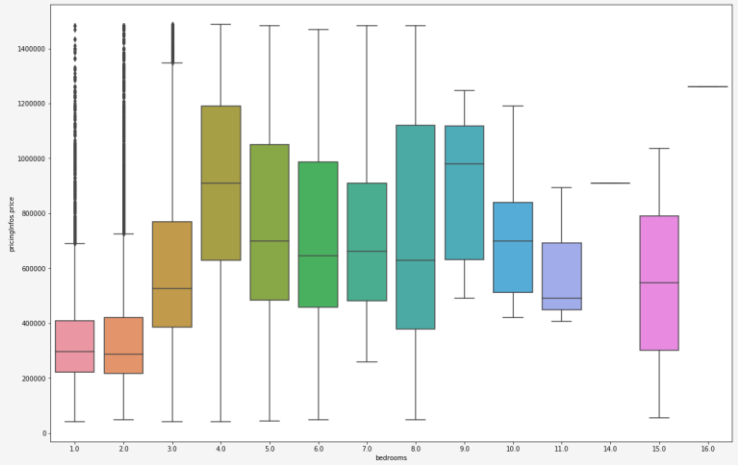


Diagrama de caixa: Price x Suites

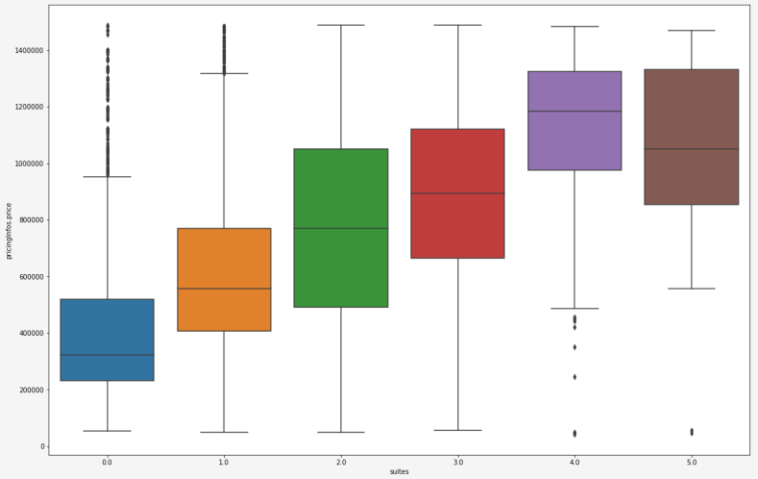
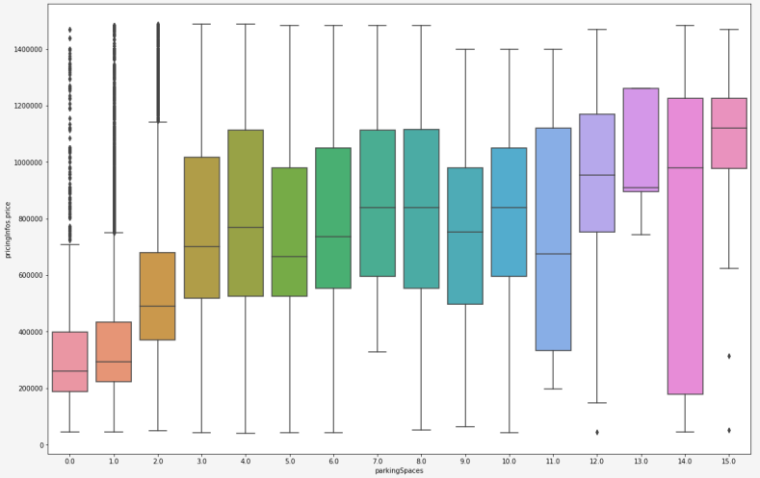


Diagrama de caixa: Price x ParkingSpaces



Análise: A mesma análise foi aplicada para as variáveis **Bedrooms, Suites e ParkingSpaces**, resultando nos diagramas sem outliers acima. Nestes diagramas pode-se observar que a média dos preços cresce de forma mais linear para a variável **suítes** do que as outras confirmando a análise do mapa de calor anterior.

“Price” em relação a MonthlyCondoFee

Gráfico de dispersão dados bruto: Price x MonthlyCondoFee

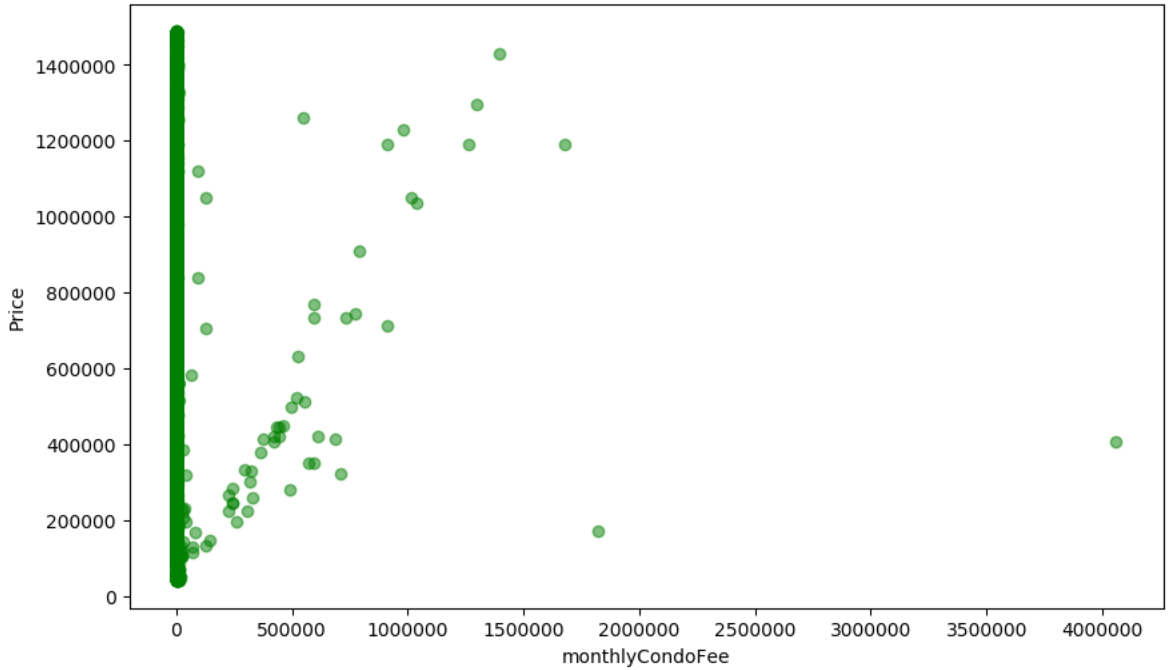
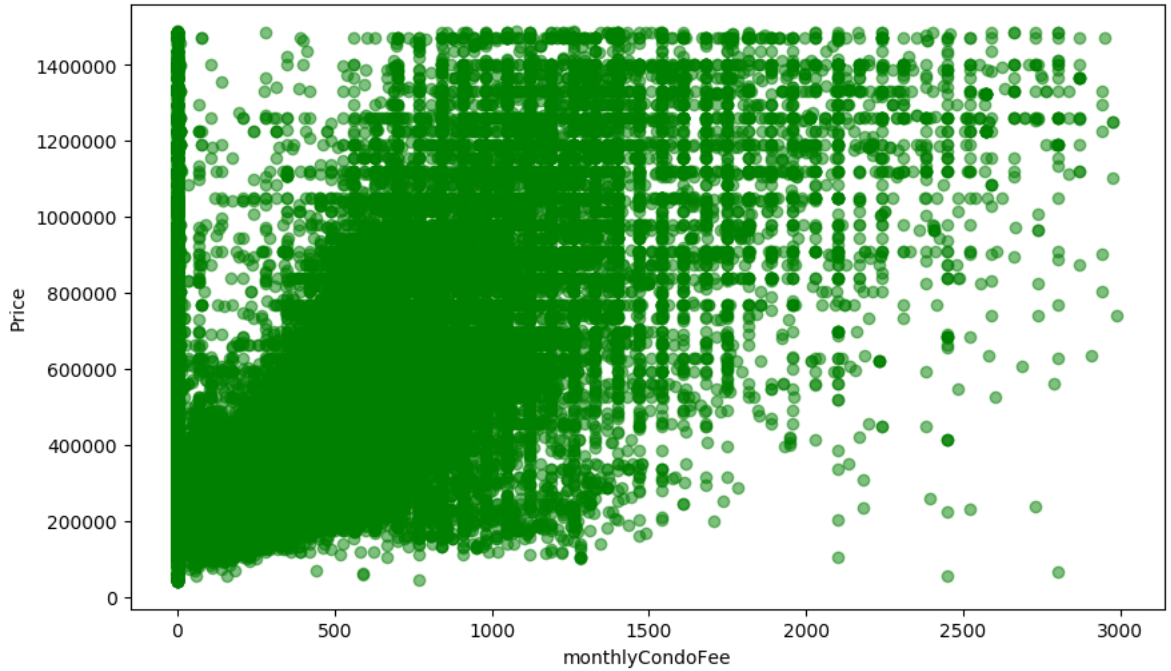


Gráfico de dispersão dados sem outliers: Price x MonthlyCondoFee



Análise: Pode-se observar que sem outliers conforme aumenta o preço do condomínio cresce o valor do imóvel.

“Price” em relação a TotalAreas, UsableAreas e YearlyIptu

DADOS SEM OUTLIERS

Gráfico de dispersão: Price x TotalAreas

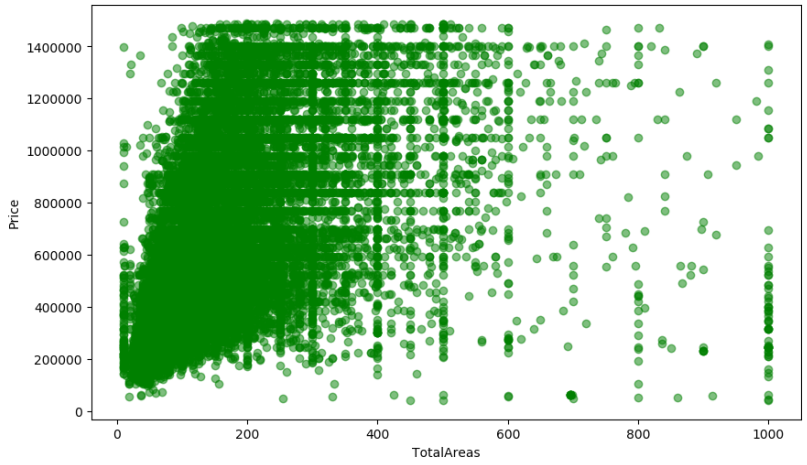


Gráfico de dispersão: Price x UsableAreas

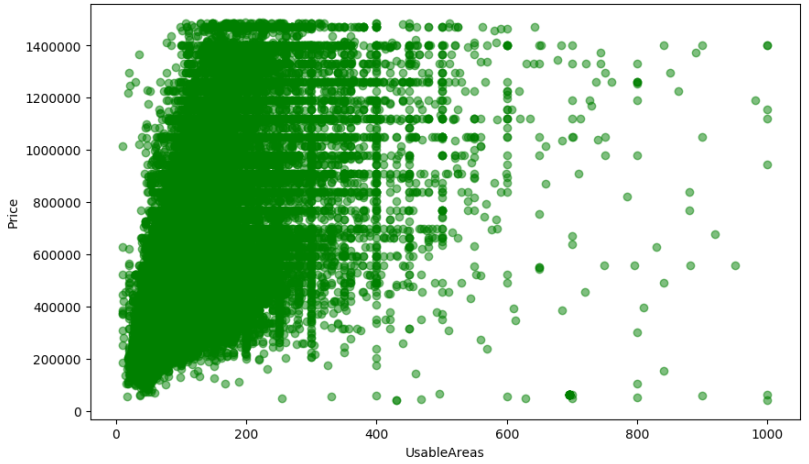
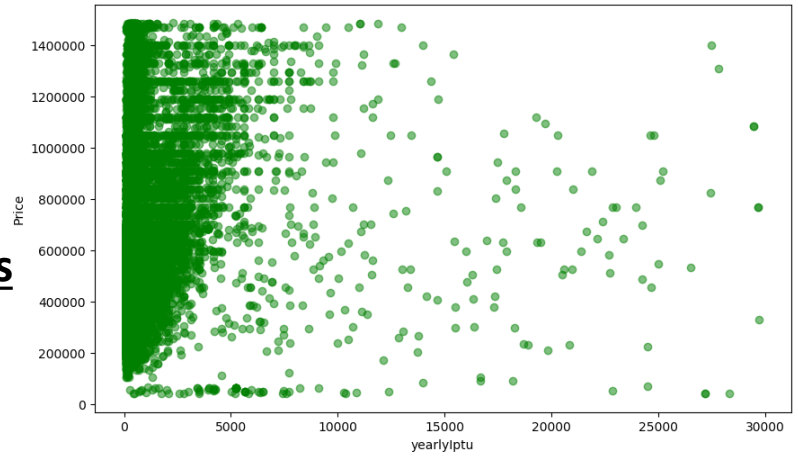


Gráfico de dispersão: Price x YearlyIptu



Análise: A mesma análise do slide anterior foi aplicada para as variáveis **TotalAreas**, **UsableAreas** e **YearlyIptu**, resultando nos gráficos sem outliers acima. Pelos gráficos pode-se observar a boa relação entre as áreas e Price.

Dados faltando

Após a eliminação dos outliers os dados que faltavam foram completados com as médias por bairro e por tipo de imóvel.

Mapa de calor dados sem outliers

Observe que após o tratamento de outliers o mapa de calor mudou. Agora, o preço dos imóveis se correlaciona melhor com o número de banheiros, tamanho do imóvel e condomínio.

Conclusão


Para estimar o preço dos imóveis os 3 melhores campos são: número de banheiros, tamanho do imóvel e condomínio.
zap

APLICAÇÃO DO MODELO DE MACHINE LEARNING

A escolha do modelo de previsão

1-Hipótese: Os dados não apresentam uma relação linear tão intensa pois existem vários bairros. Assim o modelo de machine learning deverá ser capaz de analisar os dados por bairros e aplicar uma regressão.

2-Hipótese: O modelo de **Random Forest regressor** foi escolhido por ser um algoritimo poderoso e atende a hipótese 1.

Tratamento das variáveis categóricas

As variáveis categóricas como Bairros, Nome das Ruas e Tipo do imóvel foram transformadas em variáveis numéricas para que o algoritmo possa trabalhar .

RESPOSTA DAS PEGUNTAS

Questão 1

Pergunta: Você utilizaria a métrica escolhida para seleção de modelo também para comunicar os resultados para usuários e stakeholders internos? Em caso negativo, qual outra métrica você utilizaria nesse caso?

Resposta: Sim, pois a métrica utilizada apresentou uma assertividade da previsão em torno de 70%

Questão 2

Pergunta: Em quais bairros ou em quais faixas de preço o seu modelo performa melhor?

Resposta: O modelo foi testado utilizando apenas bairros com dados acima de 100 e observou-se uma melhora no número de acertos da previsão. Assim, o modelo performa melhor para os bairros que contenham mais dados como **Santana e Pinheiros.**

Questão 3

Pergunta: Se você tivesse que estimar o valor dos imóveis com apenas 3 campos, quais seriam eles?

Resposta: Número de banheiros, tamanho do imóvel e condomínio.

Questão 4

Pergunta: Como você vislumbra colocar a sua solução em produção?

Resposta: Com o grande banco de dados sobre imóveis que Zap apresenta é possível com a modelagem criada neste projeto desenvolver uma solução para que empresas ou profissionais do setor imobiliário possam estimar os preços de imóveis. Como por exemplo, construtoras poderiam utilizar o serviço para estimar qual seria o preço de venda do imóvel após a construção e analisar os ganhos ou um perito poderia utilizar a ferramenta para auxiliar no desenvolvimento do seu laudo.

OBRIGADO