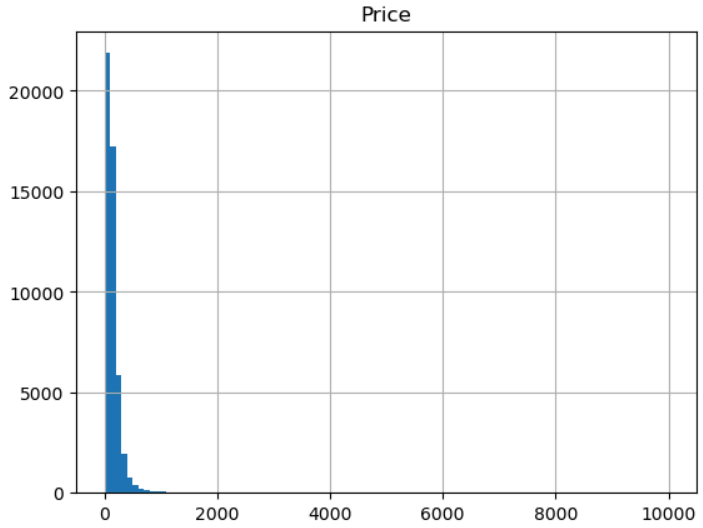
Relatório Desafio Data Science Indicium

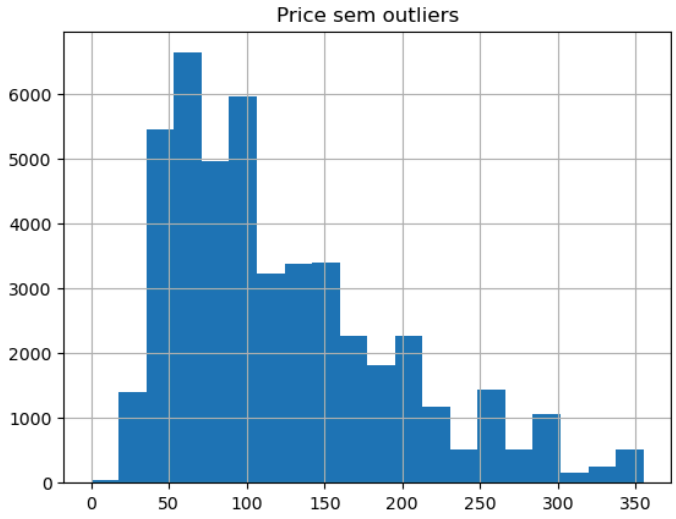
Mateus Moro Torres

1- Análise Exploratória (EDA)

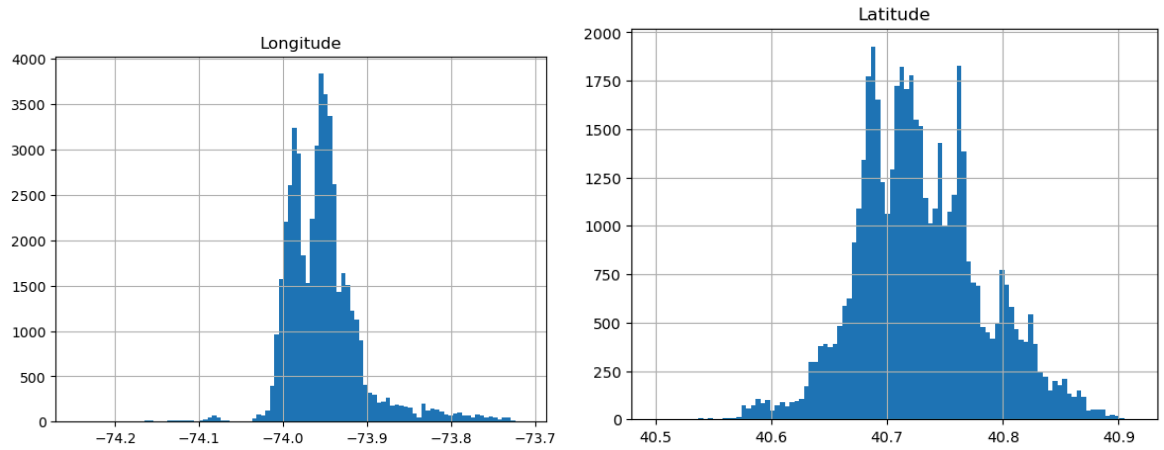
A variável preço tem seus valores entre 0 e 10000, e boa maioria concentrados entre 0 e 350.



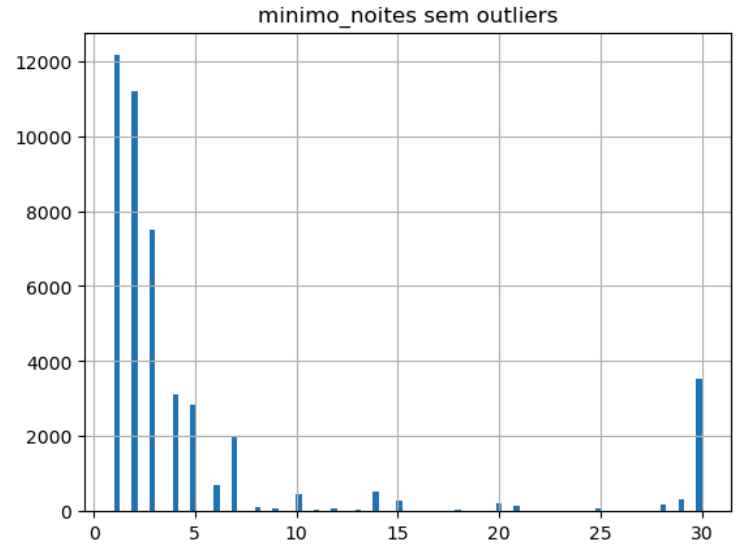
Retirando-se os outliers, considerado aqui como preços até o percentil P99, tem-se:



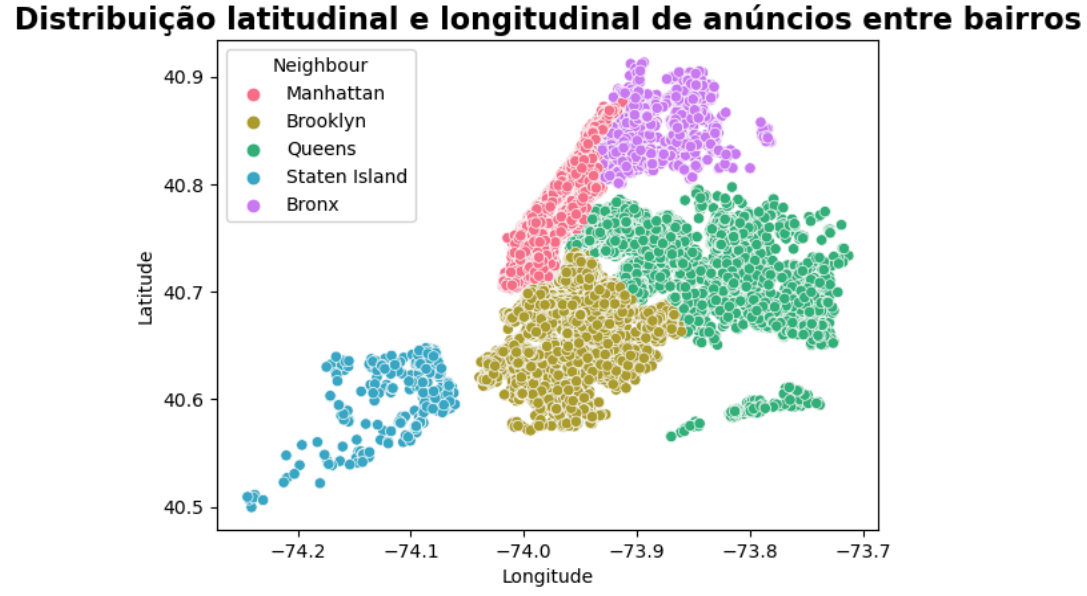
Distribuição da latitude e longitude:

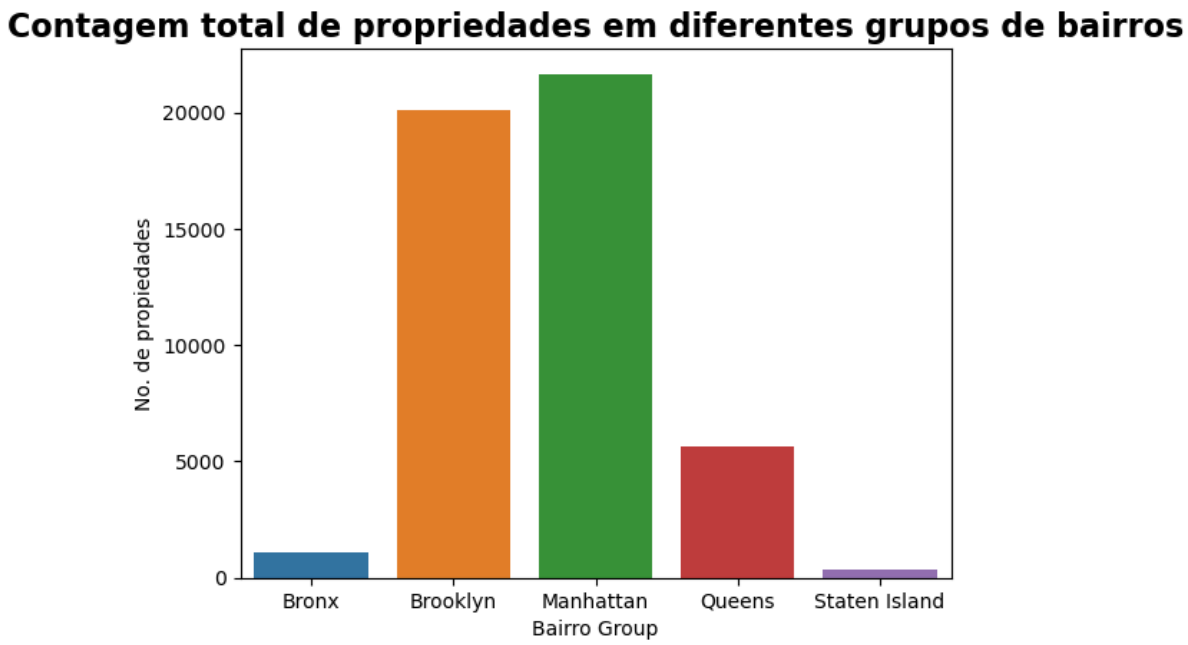


Mínimo de noites sem os outliers, já que tem valores extremos maiores que 365 dias.

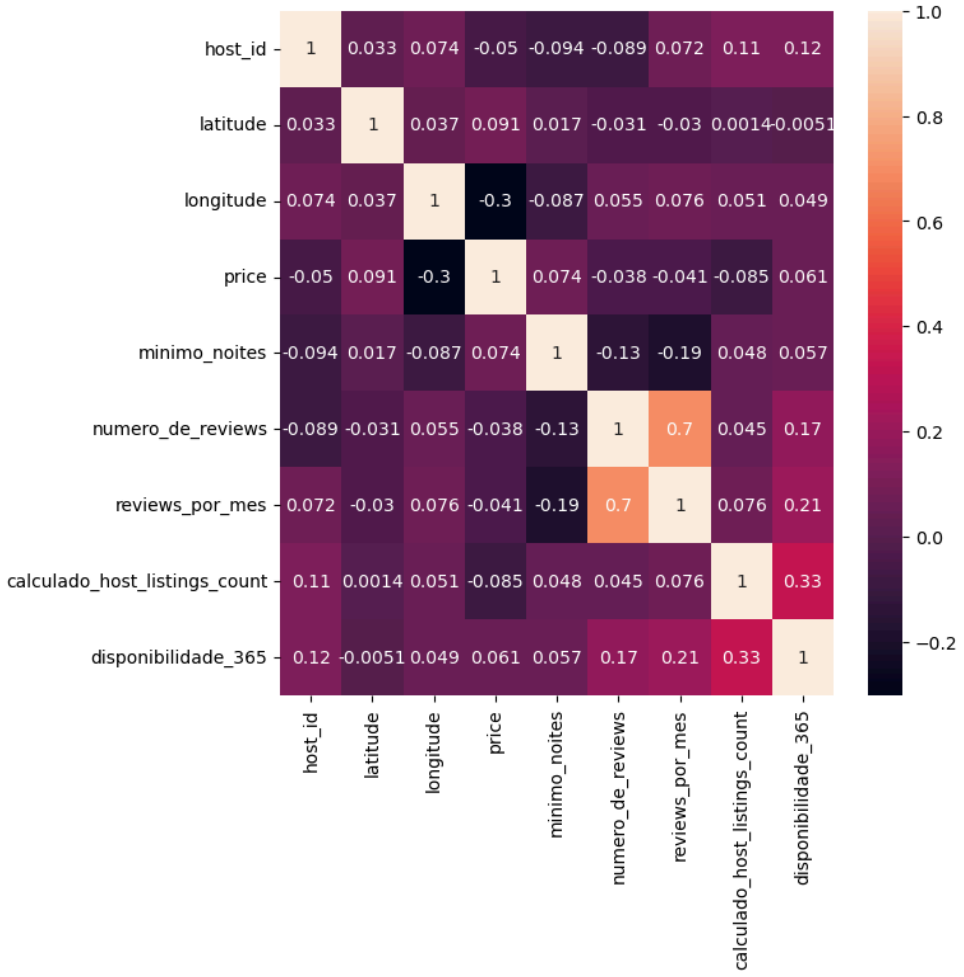


Distribuição Geográfica do Airbnbs





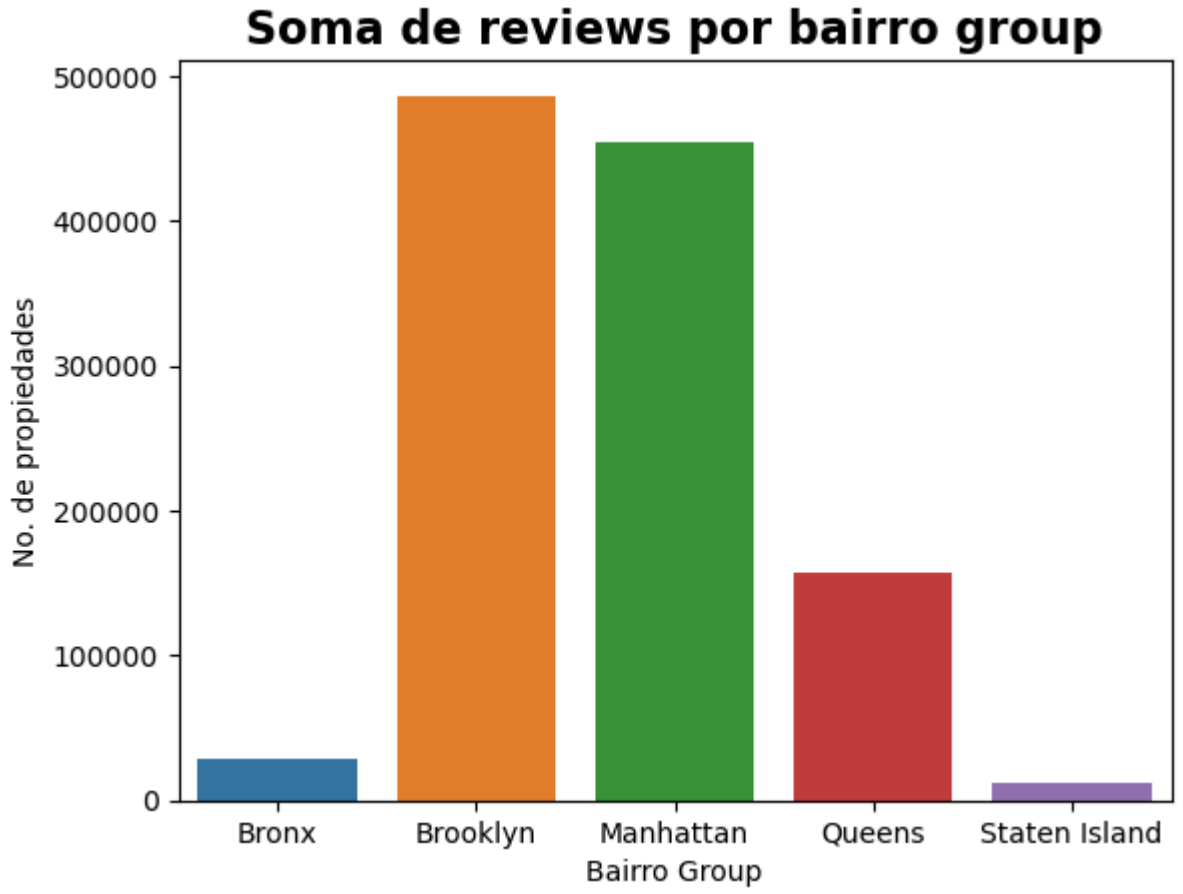
Correlação das variáveis:



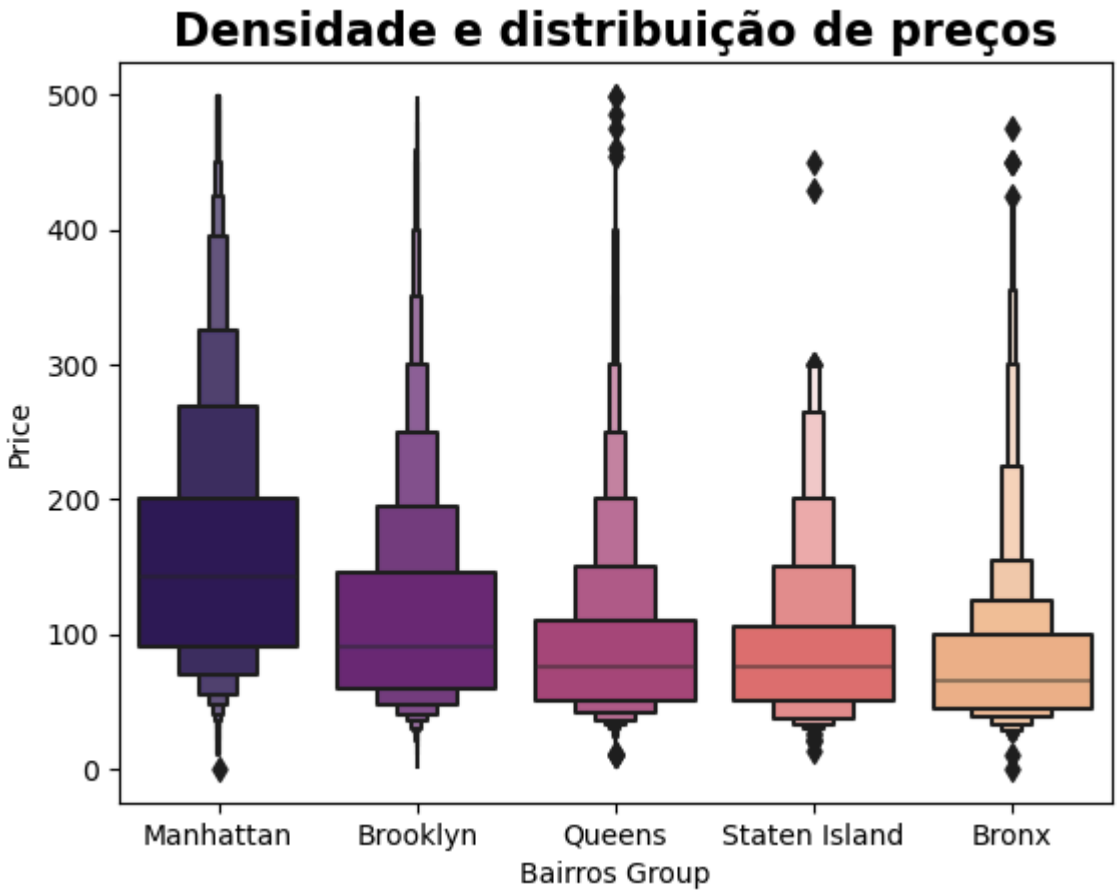
- Percebe-se que apartamentos com mais reviews tem mais reviews por mês, o que é de certa forma esperado.

- Apartamentos com maior disponibilidade tem um relação positiva com número de apartamentos listados na plataforma, o que mostra que donos com maiores disponibilidade de dias também têm mais de um imóvel no Airbnb.
- Reviews por mês tem correlação negativa com mínimo de noites, mostrando que quanto maior o mínimo de noites, menor reviews por mês.
- Preço e latitude tem correlação negativa de 0.3, o que mostra que apartamentos mais caros estão mais próximos do centro.

Soma de reviews por grupo de bairro mostra que o Brooklyn é um bairro onde se tem mais reviews de Airbnbs.

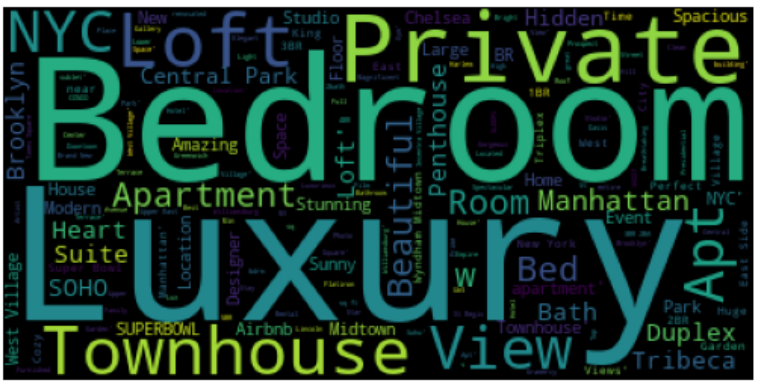


Da densidade e distribuição de preços abaixo tem-se que Manhattan tem a maior média de preços, seguido do Brooklyn.



2- Perguntas

a) Dois bons lugares para investir em um apartamento seriam os grupos de bairro Manhattan e Brooklyn, já que têm as maiores médias de preços e também o maior números de reviews. Assim sendo, os preços são elevados e tem bastante demanda.
b) Pela matrix de correlação, ambas as métricas de mínimos de noites e disponibilidade ao longo do ano tem pouca influência nos preços.
c) A palavra “Luxury”, “Private” aparece bastante para apartamentos mais caros, como mostra a wordcloud abaixo.



3- A previsão dos dados foi feita usando um modelo de regressão linear múltipla, já que se trata de um problema de regressão para estimar o preço dos Airbnbs com base em algumas variáveis.

Como somente algumas variáveis fazem sentido para explicar os preços, foram mantidas somente as seguintes:

- Bairro_group
- Longitude
- room_type
- minimo_noites
- Calculado_host_listings_count
- disponibilidade_365

Como o Bairro_group e room_type são variáveis categóricas, foi feito um “one-hot enconding” dessas colunas, de modo que fossem tratadas como vetores binários, onde cada coluna representa um tipo de categoria.

O modelo de regressão linear múltipla é fácil de se usar e captura relações lineares entre as variáveis. Os pontos negativos são que pode não ser adequado para relações não lineares, pode ser sensível a outliers e assume que os resíduos seguem uma distribuição normal.

Foi escolhida a medida de performance do coeficiente de determinação(R^2) já que representa a proporção da variabilidade da variável dependente é explicada pelas variáveis independentes, ou seja, indica o quanto o erro na previsão do preço é devido a fatores não considerados no modelo. Outra medida escolhida foi o RMSE(Root Square Mean Error) que é uma medida do erro absoluto, é a média entre os valores reais e os valores previstos pelo modelo.

4 - Para o apartamento em questão o preço sugerido seria de 122.41.