Desafio Cientista de Dados

Introdução

O presente relatório aborda a construção de uma análise sobre os preços de aluguéis na cidade de Nova York, a fim de tornar possível uma precificação mais assertiva. O projeto completo foi feito em duas etapas, em que a primeira se refere à análise exploratória dos dados e aos testes de hipóteses de negócio, enquanto a segunda corresponde a modelagem preditiva para otimização da estratégia de precificação de aluguéis. É importante salientar que neste relatório serão mostradas apenas as descobertas da primeira etapa do projeto.

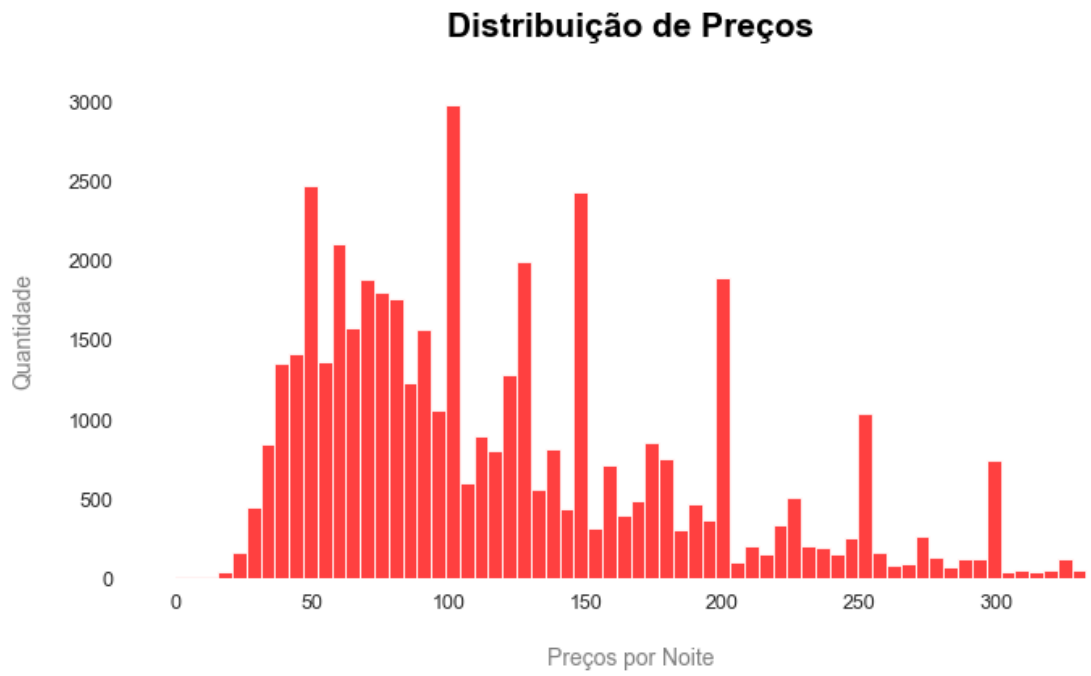
Dados

Neste projeto foram utilizadas duas bases de dados:

- teste_indicium_precificacao: Base de dados fornecida que contém informações como bairro do imóvel, tipo de apartamento e preço do aluguel. Essa base de dados possui 48.894 observações e 15 colunas.
- NY-House-Dataset: Base de dados baixada na plataforma Kaggle que contém informações adicionais sobre preços de imóveis na cidade de Nova York. Essa base de dados possui 4.801 observações e 17 colunas.

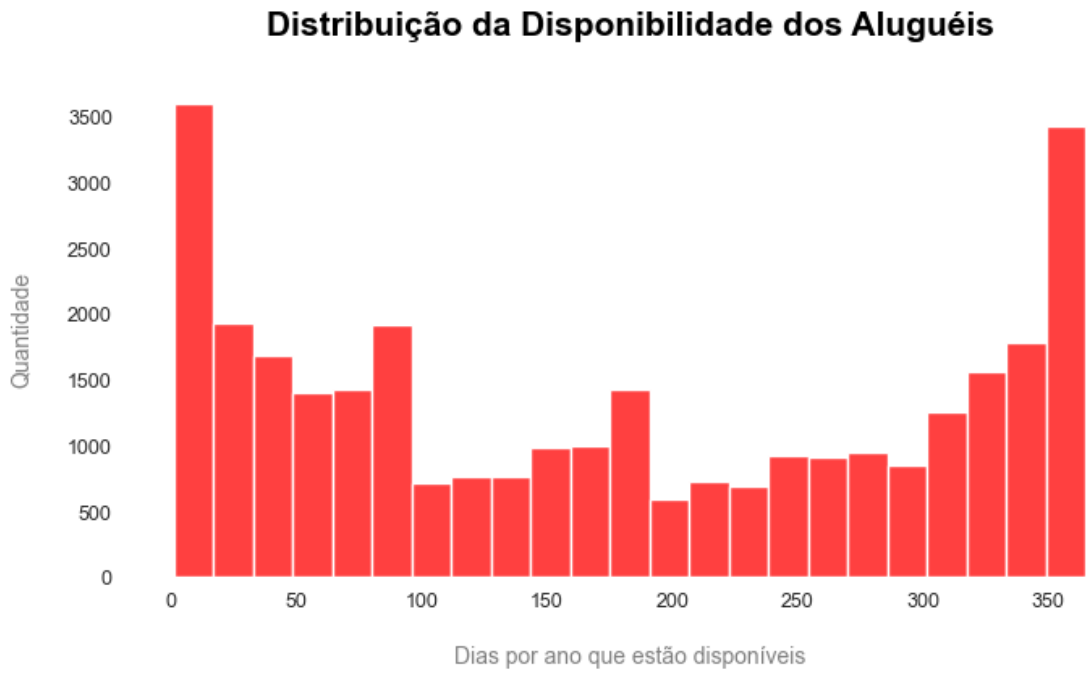
Análise Exploratória de Dados

Em primeiro lugar, a fim de analisar a distribuição de preços e perceber quais modelos seriam mais adaptáveis aos dados, foram utilizados alguns histogramas:

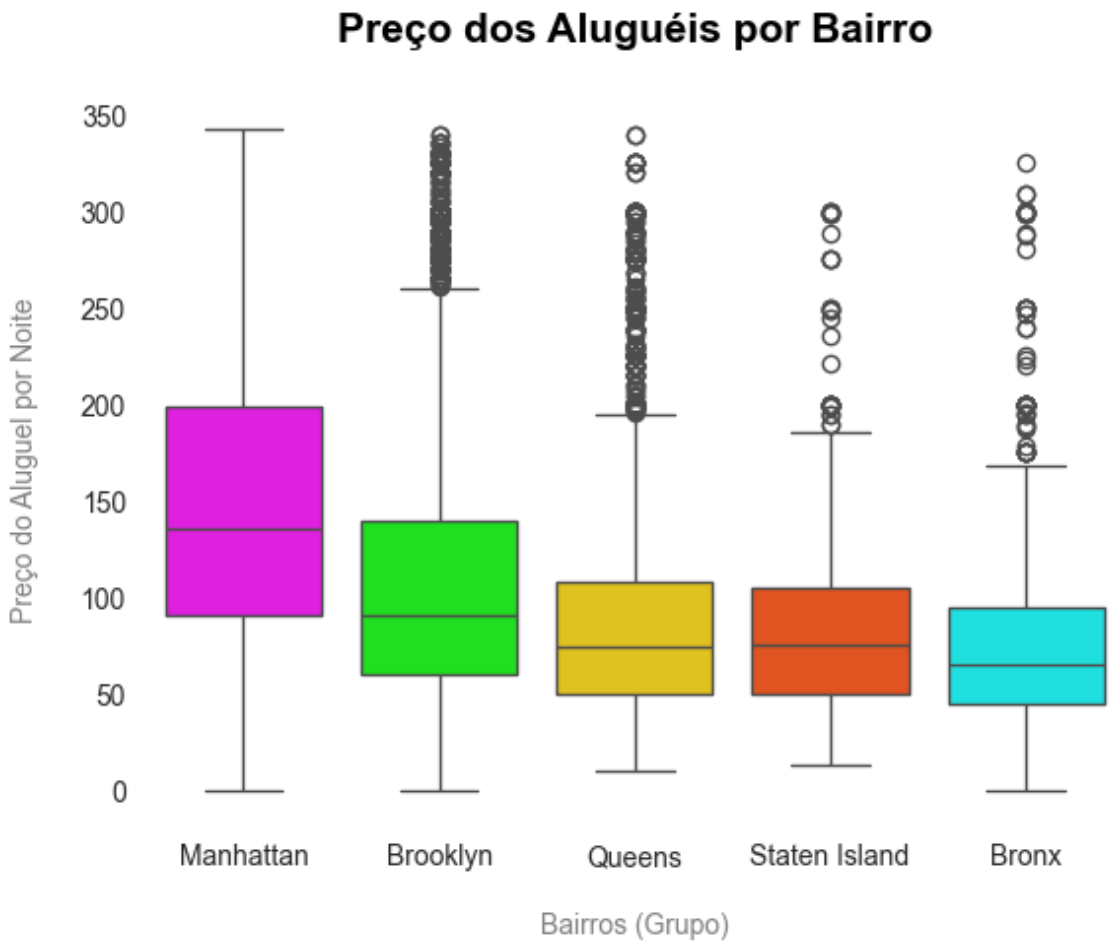


Primeiramente, é interessante mencionar que este gráfico desconsiderou preços categorizados como *Outliers* para que a visualização não fosse comprometida. Dessa forma, como é possível perceber, a distribuição de preços de aluguel por noite revela que existe uma grande concentração dos aluguéis na faixa de 50 a 150 dólares.

Investigando a relação entre a disponibilidade dos aluguéis ao longo do ano, percebe-se o nível que dados extremos apresentam na base de dados. Observando o gráfico abaixo, é possível identificar que existem muitos apartamentos com disponibilidade ampla ao longo do ciclo anual, bem como apartamentos com disponibilidade ínfima.



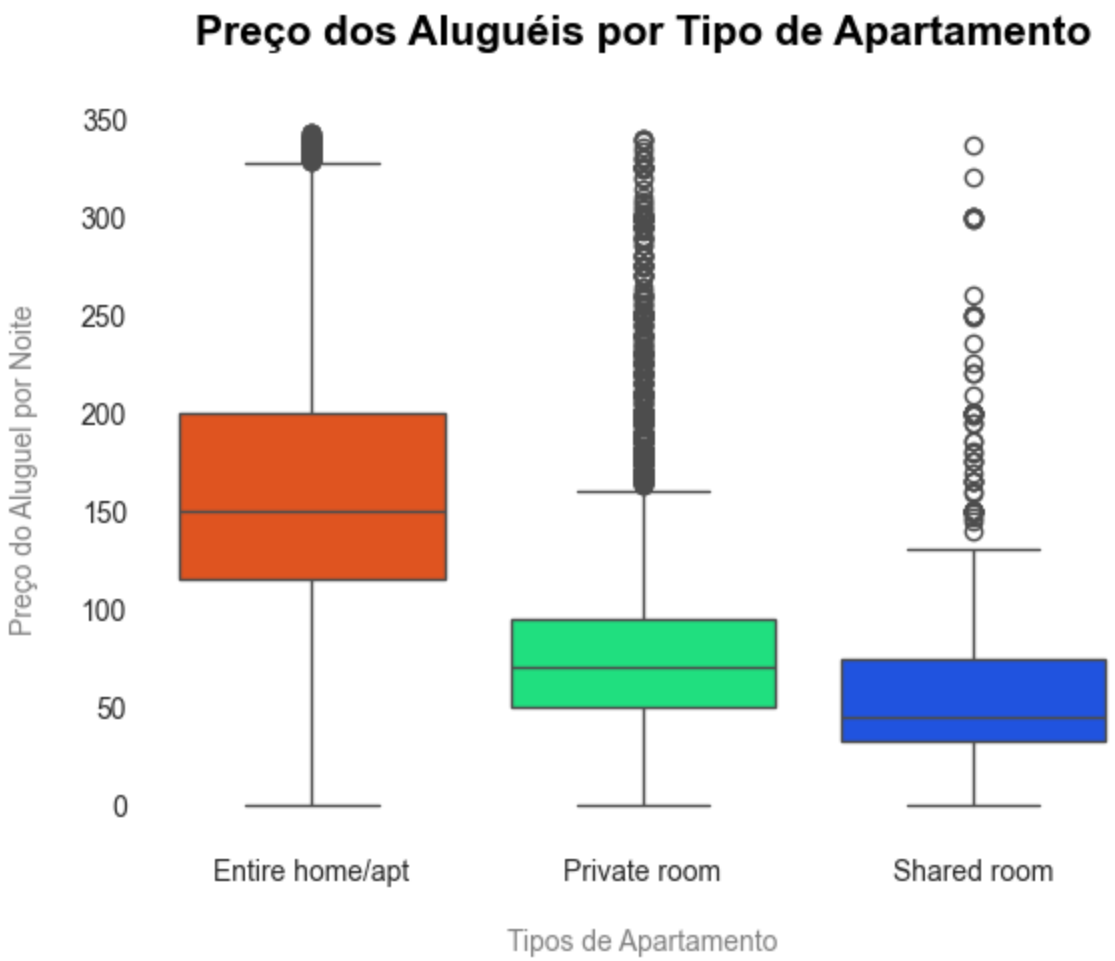
Partindo para a análise exploratória bivariada foi possível apontar algumas **hipóteses de negócio** envolvendo o preço dos aluguéis e determinadas variáveis. **A primeira hipótese de negócio testada referiu-se à localidade do imóvel exercer influência sobre o preço de seu aluguel**. A fonte primária dessa hipótese foi a visualização gráfica do boxplot que relaciona o preço do aluguel por noite de acordo com cada bairro. A partir da análise gráfica, foi possível perceber que existem diferenças entre os preços nos 5 bairros.



Como é possível perceber, geralmente os aluguéis de apartamentos em Manhattan apresentam preços maiores se comparados aos demais bairros. Observando a mediana de cada boxplot, percebe-se que os aluguéis do Bronx aparentam ser os mais baratos.

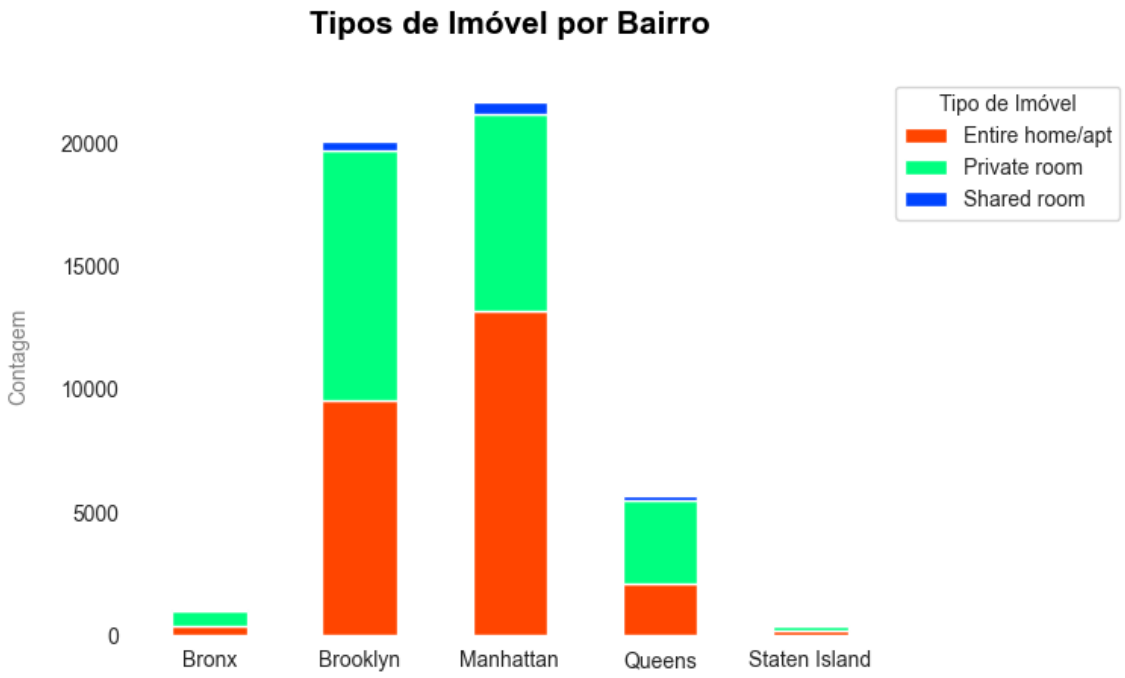
No entanto, a fim de consolidar a hipótese de negócio formulada, foram utilizados testes estatísticos (qui-quadrado) para verificar se a localidade afeta o preço dos aluguéis. O teste aplicado indicou que, de fato, a **localidade possui influência sobre o preço do aluguel**. A partir disso, foi utilizada a metodologia ANACOR para verificar a **segunda hipótese de negócio, que revelou que os imóveis de Manhattan, são, efetivamente, relacionados a aluguéis mais caros.**

A terceira hipótese de negócio analisada referiu-se à influência do tipo de apartamento sobre seu preço, e, a partir da utilização do teste qui-quadrado, foi possível concluir que o tipo de apartamento de fato influencia o preço dos aluguéis.



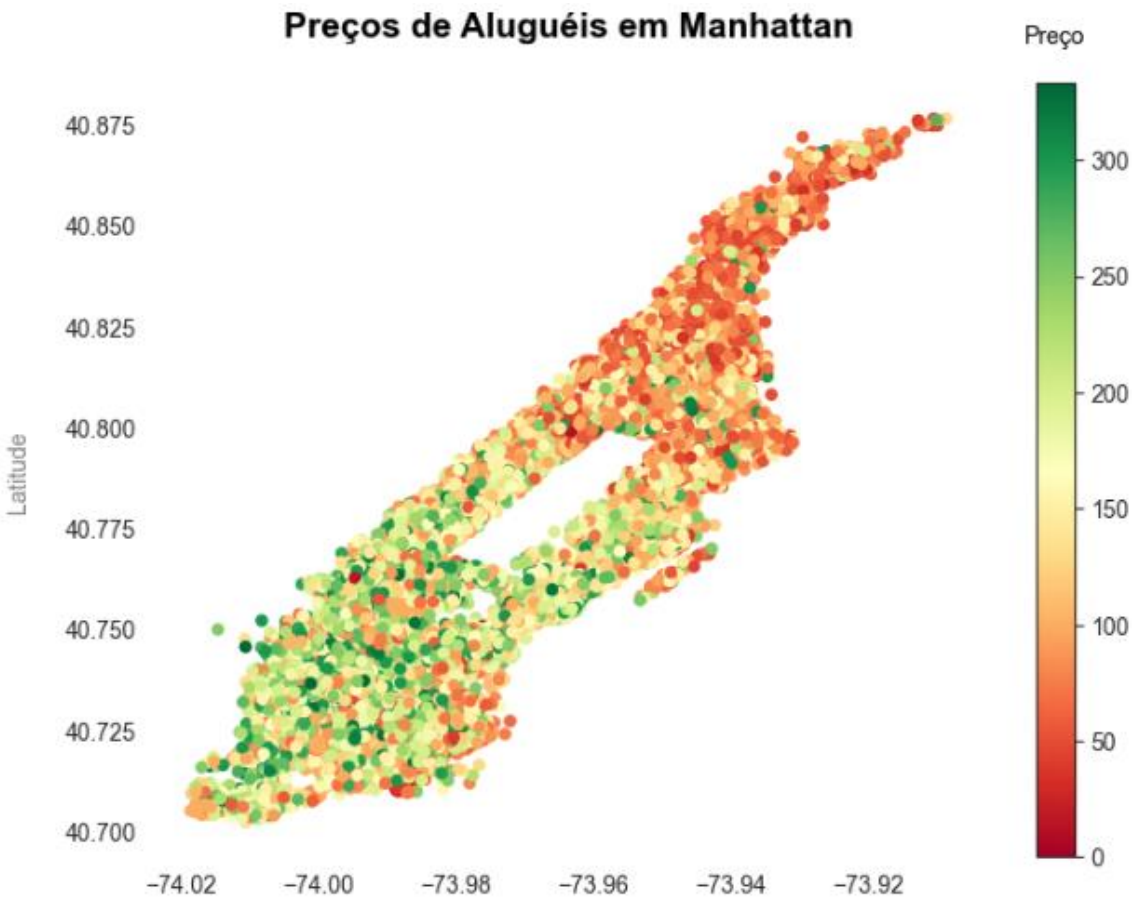
Diante dessa informação, a metodologia ANACOR foi aplicada mais uma vez para verificar quais categorias da variável tipo de apartamento estavam relacionadas aos preços de aluguéis mais caros. Portanto, a **quarta hipótese de negócio analisada referiu-se ao questionamento da relação dos apartamentos do tipo Entire Home com preços mais altos.** Conforme demonstrado pela metodologia utilizada (nos Notebooks deste projeto), verificou-se que **a relação dos apartamentos do tipo Entire Home com preços de aluguéis mais altos é estatisticamente significante**.

Em seguida, foi analisada a composição dos bairros, em termos de tipos de apartamento:



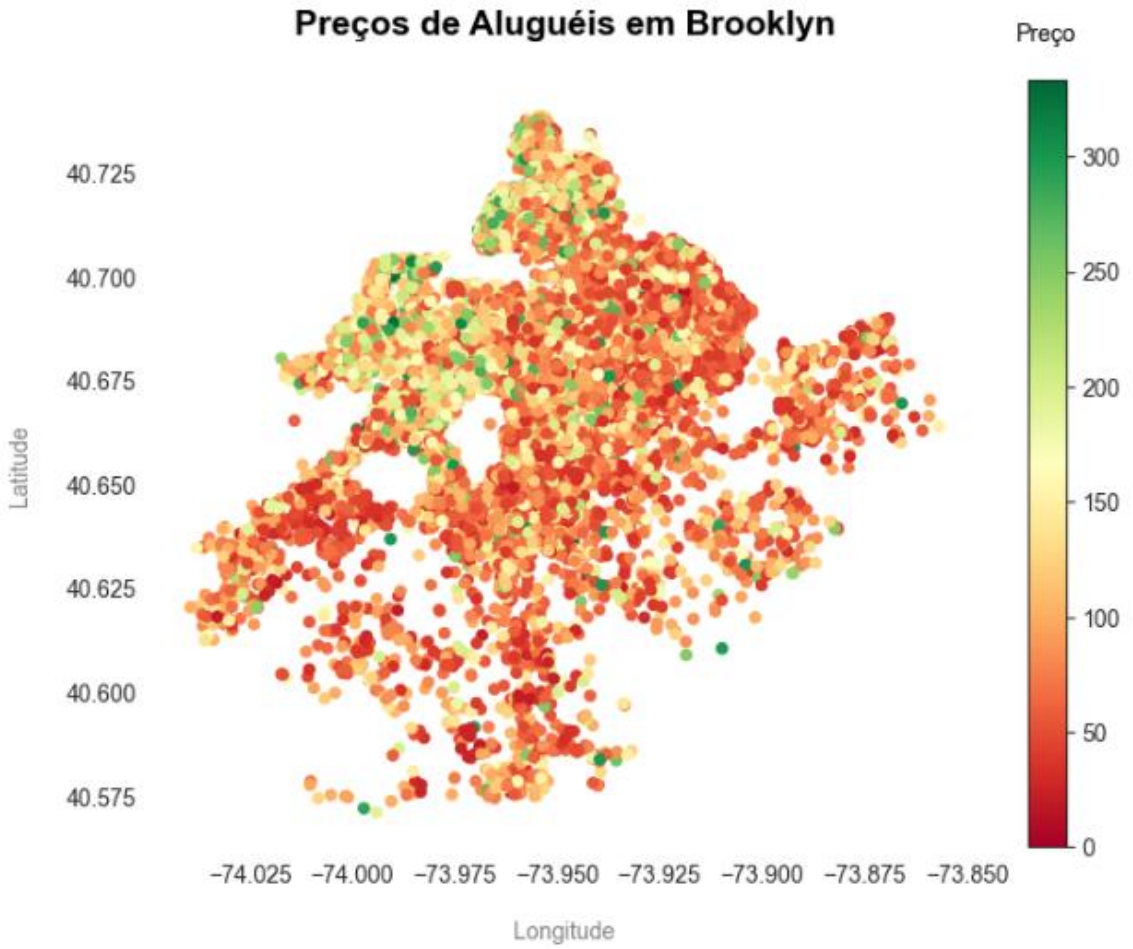
Como é possível observar, os bairros de Manhattan e Brooklyn (que possuem os aluguéis mais caros) são compostos, em grande parte, pelos apartamentos do tipo Entire Home, que, conforme mencionados anteriormente, são os apartamentos relacionados estatisticamente a aluguéis mais elevados.

Levando em consideração dados espaciais disponibilizados, foi possível analisar mais a fundo a distribuição dos preços de aluguéis dentro do bairro de Manhattan.



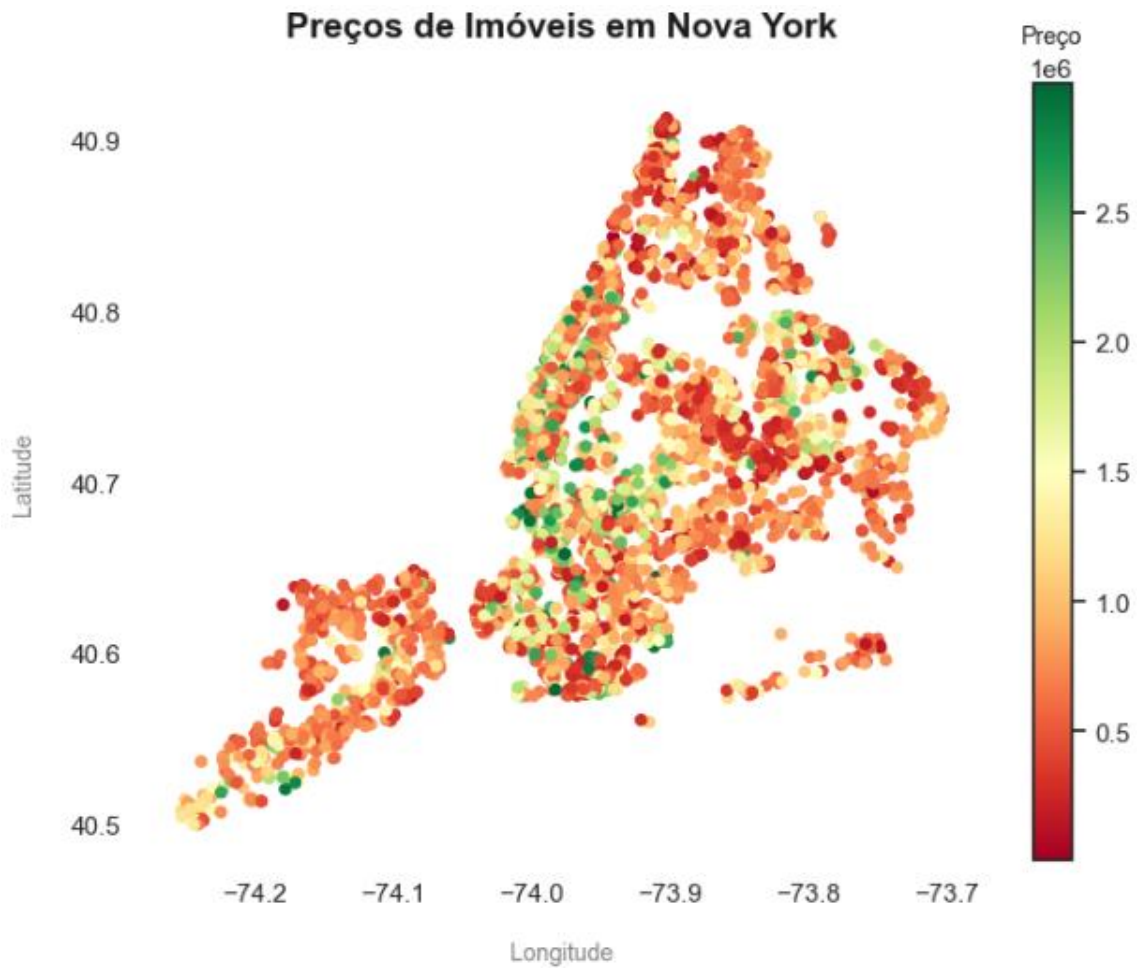
De acordo com o gráfico de dispersão analisado, foi possível concluir que existe uma alta concentração de preços de aluguéis mais caros na região sudoeste de Manhattan. Isso pode ser explicado em razão das atrações turísticas que existem nessa região, como a *Times Square* e a proximidade da *Statue of Freedom*. É interessante perceber também que o espaço em branco no meio do bairro de Manhattan refere-se ao *Central Park*, havendo alguns apartamentos de aluguéis mais caros em seus arredores.

Já no que se refere aos apartamentos disponíveis para aluguel no Brooklyn, o seguinte gráfico foi analisado:



Neste caso, percebeu-se que embora não haja um padrão de concentração como em Manhattan, a região norte do Brooklyn apresenta leve concentração de aluguéis mais caros (o que pode estar relacionado à sua proximidade com Manhattan).

Ao utilizar a base de dados baixada no Kaggle, percebeu-se que a análise foi enriquecida, uma vez que os dados acerca dos preços dos **imóveis** na cidade de Nova York demonstraram a mesma situação mostrada acima.



A concentração de preços mais caros (lembrando que a escala está em milhões de dólares, uma vez que a variável plotada refere-se ao preço dos imóveis) ocorre no centro de Nova York, que corresponde a localização de Manhattan.

Respondendo Perguntas de Negócio

Além da análise exploratória de dados, foram respondidas determinadas perguntas de negócio propostas no desafio:

As respostas para cada pergunta serão resumidas nos tópicos abaixo:

a) Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

A fim de responder essa pergunta é interessante relembrar os resultados obtidos na verificação das Hipóteses de Negócio 1 e 2.
Como foi demonstrado anteriormente, existe uma relação significativa entre os preços dos aluguéis por noite com a localidade do imóvel. De forma mais profunda, foi constatado (por meio da utilização da análise de correspondência - ANACOR) que os imóveis localizados em Manhattan geralmente apresentam valores mais altos para o aluguel. Além disso, é importante destacar que, com a utilização do gráfico de dispersão baseado nos dados de latitude e longitude, foi possível observar uma concentração de apartamentos com aluguel mais caro na região sudoeste de Manhattan, possivelmente em virtude dos pontos turísticos dessa localidade, como a Times Square.
Também é válido levar em consideração que os preços dos aluguéis por noite dos imóveis do tipo Entire Home são os mais caros dentre os outros tipos de imóveis.

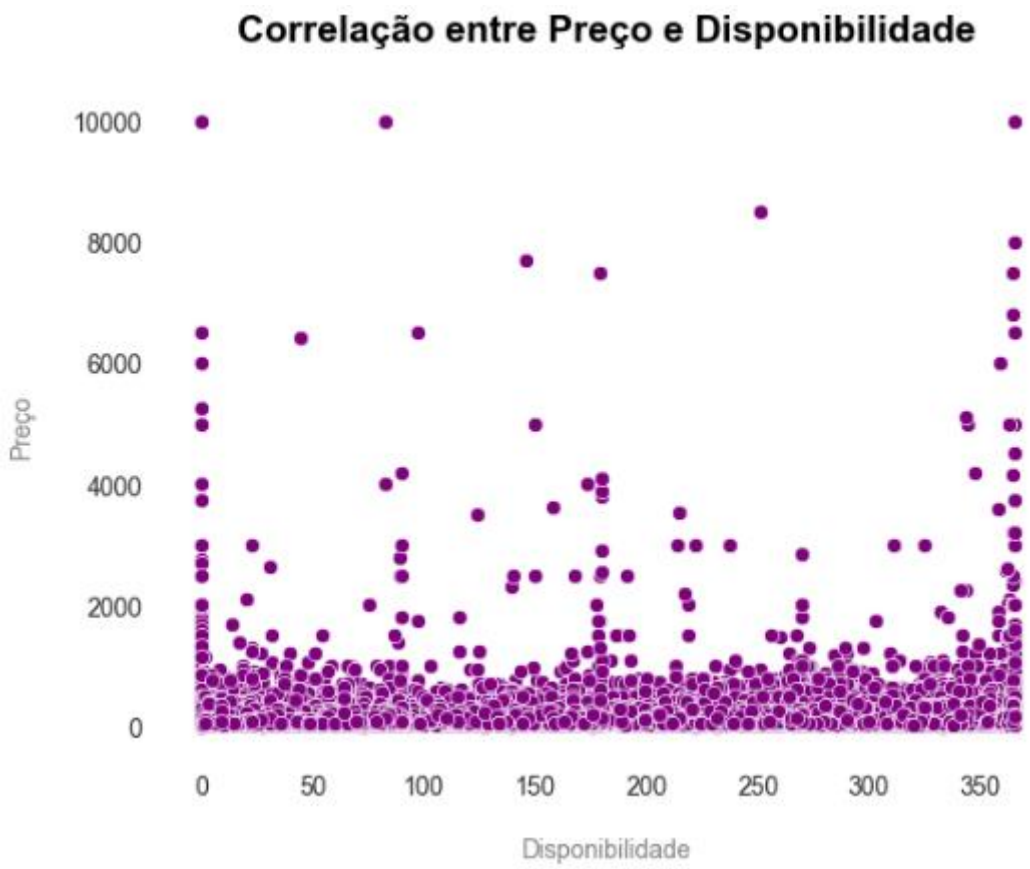
Tendo isso em mente, é possível concluir que a compra de um apartamento é mais vantajosa na **região sudoeste de Manhattan**, onde os preços dos aluguéis são mais **altos**. Além disso, se essa pessoa deseja ampliar ao máximo o rendimento de seu investimento, é **interessante considerar a compra de um apartamento do tipo Entire Home/Apt**.

b) O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

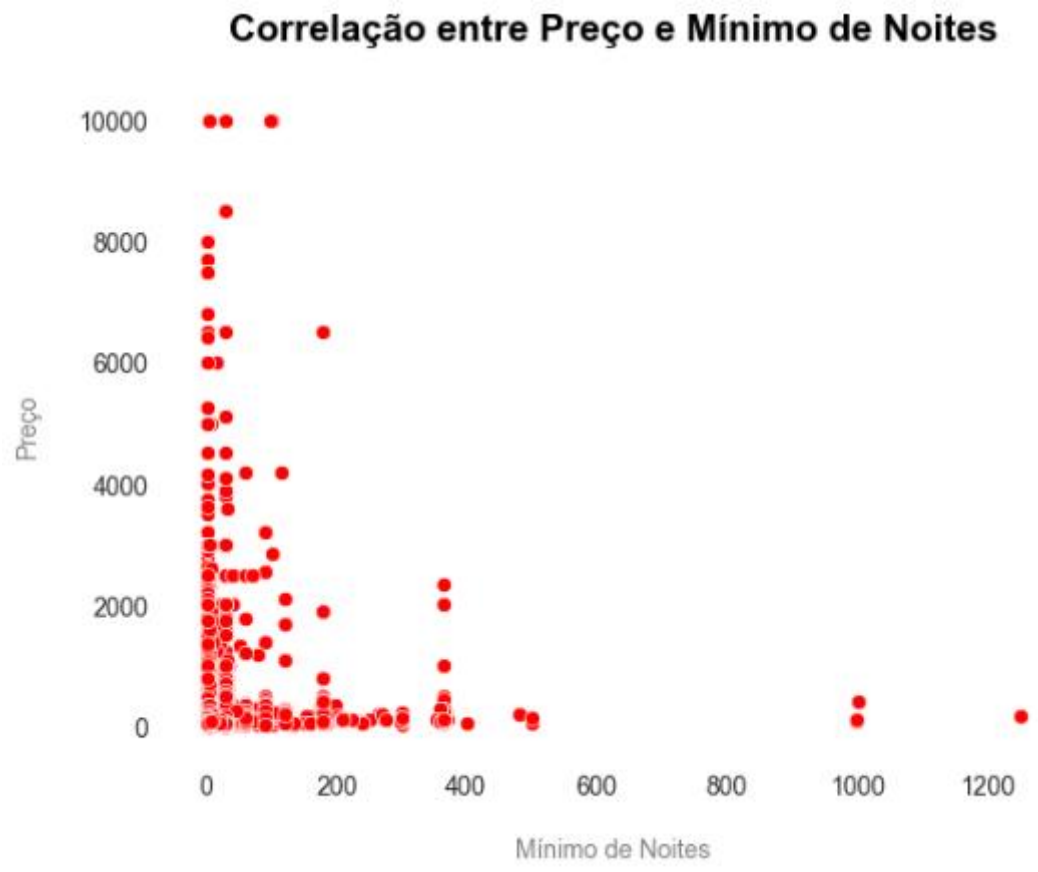
Para responder essa pergunta foi realizado um teste de correlação de Pearson, a fim de verificar a relação entre tais variáveis e o preço dos aluguéis.

O teste indicou que o mínimo de noites e a disponibilidade ao longo do ano interferem no preço, porém, a relação **linear** entre mínimo de noites e preço, bem como entre disponibilidade ao longo do ano e preço são fracas.

Os gráficos de dispersão foram plotados para verificar essa relação **linear** (teste de correlação de Pearson só testa relações lineares), e foi possível perceber que a correlação **linear** entre preço e disponibilidade ao longo do ano foi de 0.08.



Já a correlação entre preço e mínimo de noites foi ainda menor (0.04).



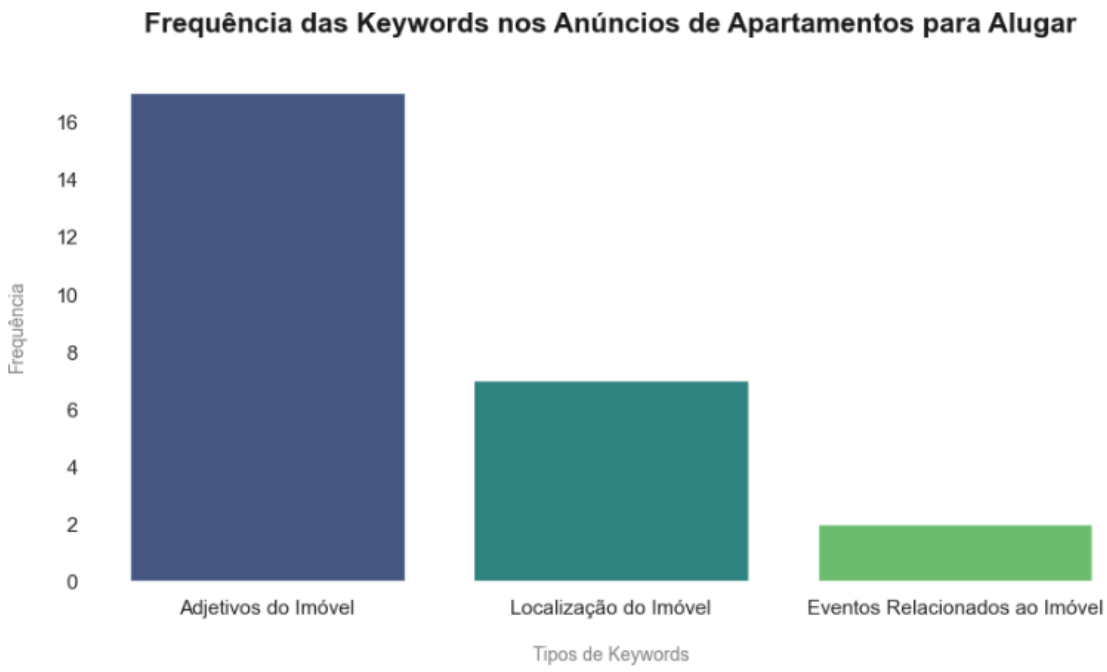
c) Existe algum padrão no texto do nome do local para lugares de mais alto valor?

Para responder a essa pergunta de negócio, foi necessário filtrar os dados para visualizar apenas aqueles com preços muito altos. Nesse sentido, observando os nomes dos apartamentos cujo aluguel está listado acima dos 2000 dólares, percebe-se a presença de um certo padrão de nomes. Muitos desses imóveis apresentam descrições envolvendo bairros prestigiados de Nova York, como "Nolita/Soho", "Upper East", "Greenwich Village" e "TriBeCa". Além disso, também é possível identificar que muitos desses nomes apresentam adjetivos chamativos, como "Beautiful", "Luxury", "Stunning" e "Architecturally Stunning". Outro nome que apareceu dentre os aluguéis mais caros é relacionado a eventos esportivos, como é o caso de "SuperBowl".

id	price	nome
363673	3000	Beautiful 3 bedroom in Manhattan
826690	4000	Sunny, Family-Friendly 2 Bedroom
893413	2500	Architecturally Stunning Former Synagogue!
1448703	5000	Beautiful 1 Bedroom in Nolita/Soho
2110145	6000	UWS 1BR w/backyard + block from CP
2224896	4000	NYC SuperBowl Wk 5 Bdrs River View
2243699	5250	SuperBowl Penthouse Loft 3,000 sqft
2271504	6500	SUPER BOWL Brooklyn Duplex Apt!!
2274084	2750	3 Bedroom Apartment
2276383	2500	Penthouse with Private Rooftop for Events/Shoots
2281142	3750	Prime NYC Location for Super Bowl
2659183	2300	Luxury 5BR Townhouse, Upper East
2919330	5000	NearWilliamsburg bridge 11211 BK
2952861	4500	Photography Location
2953058	8000	Film Location
4262120	2695	Columbus Circle and Park Views
4737930	9999	Spanish Harlem Apt

Sendo assim, foi realizada uma filtragem de palavras-chave para identificar as palavras que mais apareceram entre esses nomes.

Nesse sentido, foi percebido que palavras-chave relacionadas a qualificações do imóvel (adjetivos), localização prestigiada e eventos relacionados ao imóvel foram aparecendo dentre os apartamentos de aluguéis mais caros.



Como foi discutido anteriormente, dentre a descrição dos apartamentos disponíveis para aluguel com preço mais elevado, percebe-se que sempre há ênfase em palavras que qualificam o imóvel (Adjetivos do Imóvel), bem como palavras que explicitam a localização do imóvel, sobretudo em bairros prestigiados, e que indicam a presença de eventos relacionados ao imóvel.

Portanto, conclui-se que **o padrão que existe no nome dos anúncios dos apartamentos mais caros representa a referência à qualidade, localização, ou aos eventos relacionados ao imóvel**.

Conclusões da Análise Exploratória de Dados

Conforme foi mostrado durante toda a exploração dos dados, percebe-se, de acordo com as duas bases de dados, que se uma pessoa quiser investir na compra de um imóvel para alugar, **é interessante que seja considerada a compra de um apartamento do tipo Entire Home no bairro de Manhattan, e, mais especificamente, na região sudoeste. Além disso, é interessante que o anúncio do aluguel desse apartamento contenha adjetivos do imóvel, bem como a descrição de sua localidade.**

Dessa forma, é possível tornar o investimento mais **rentável**, tendo em vista o que os dados apresentaram.