Análise exploratória - Dados de aluguéis de quartos em New York

Lucas Cruz Araújo

Contexto

Viajantes e anfitriões têm utilizado uma plataforma de hospedagem para ampliar as possibilidades de viagem e proporcionar experiências mais únicas e personalizadas ao redor do mundo. O conjunto de dados que será analisado descreve a atividade de aluguéis e métricas na cidade de Nova York.

Descrição dos dados

O conjunto de dados refere-se a informações sobre casas disponíveis para aluguel, com uma dimensionalidade de 48.894 linhas e 16 colunas. As colunas incluem:

Tabela 1 - Dicionário de dados

Coluna	Tipo de dado
Id	Int
nome	string

host_id	int
host_name	string
bairro_group	string
bairro	string
latitude	int
longitude	int
room_type	string
price	int
minimo_noites	int
numero_review	int
útima_review	date
reviews_por_mes	float
calculo_host_listing_count	int
disponibilidade_360	int

Nesse dataset existiam dados nulos estes:

Tabela 2 - Valores nulos nas colunas

Coluna	Quantidade de dados nulos
nome	16
host_name	21
ultima_review	10052
reviews_por_mês	10052

Os dados categóricos da coluna "bairro_group" apresentam os seguintes valores únicos: 'Manhattan', 'Brooklyn', 'Queens', 'Staten Island', 'Bronx'. Na coluna "room_type", os valores únicos são: 'Entire home/apt', 'Private room', 'Shared room'. Além disso, a coluna "bairro" possui 221 valores distintos.

Limpeza de dados

Primeiramente, é perceptível a falta de padronização nos nomes das colunas, o que representa uma área de melhoria. Dado que se trata de dados de uma cidade nos Estados Unidos, vamos renomear as colunas para nomes totalmente em inglês.

Tabela 3 - Novos nomes de colunas

Coluna antiga	**Nova coluna**
Id	Listing_ID
nome	Listing_Name
host_id	Host_ID
host_name	Host_Name
bairro_group	Neighborhood_Group
bairro	Neighborhood
latitude	Latitude
longitude	Longitude
room_type	Room_Type
price	Price
minimo_noites	Minimum_Nights
numero_review	Number_of_Reviews
útima_review	Last_Review_Date
reviews_por_mes	Reviews_Per_Month
calculo_host_listing_count	Host_Listings_Count
disponibilidade_360	Availability_365

Os dados que continham valores nulos nas colunas 'Listing_Name' e 'Host_Name' foram removidos. Quanto aos dados que apresentavam valores nulos nas colunas 'Last_Review_Date' e 'Reviews_Per_Month', eles foram mantidos, uma vez que a coluna 'Number_of_Reviews' tinha valor zero, o que indica que não houve revisões

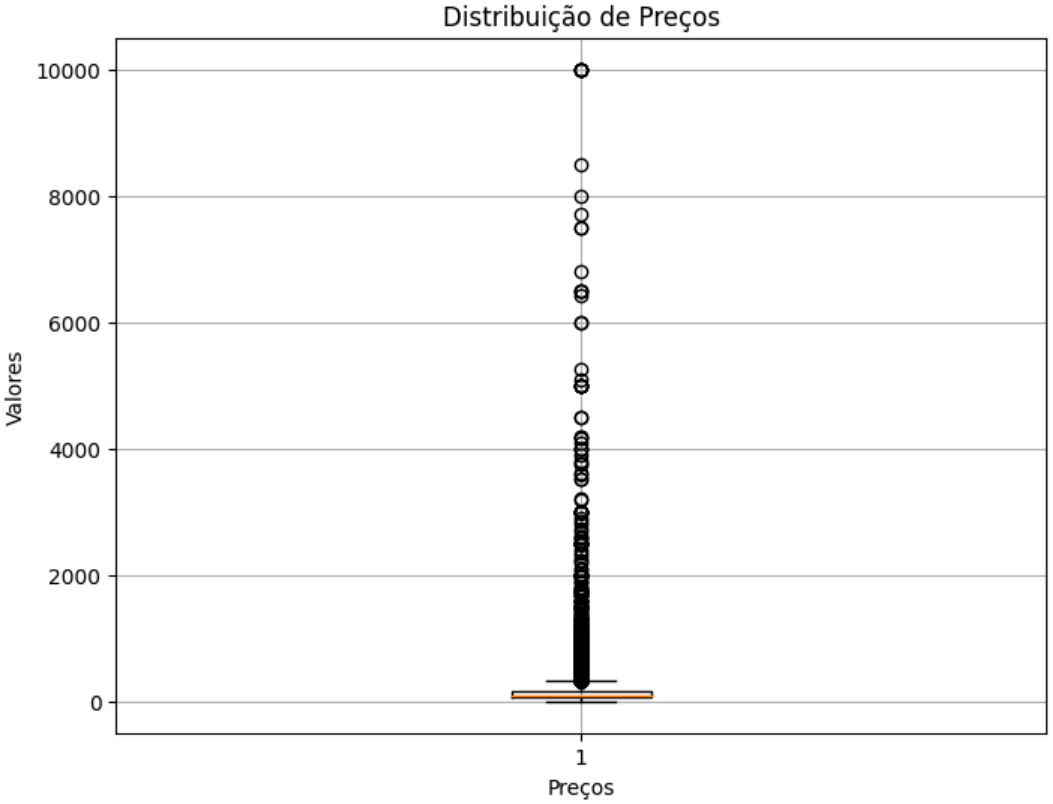
para esses registros. É importante para análises posteriores manter esses dados, então os valores nulos foram substituídos por zero.

As colunas 'Listing_ID', 'Host_ID', 'Host_Name', 'Last_Review_Date' foram removidas, pois não serão necessárias para as análises planejadas.

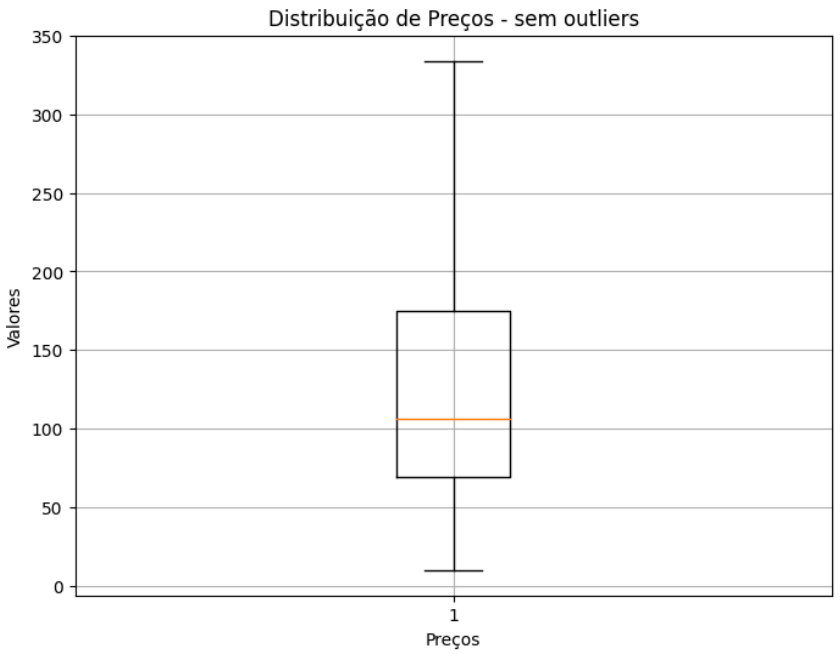
No dataset, na coluna de preços, foram removidos os valores 0, pois esses valores de aluguel são irrealistas.

Análise

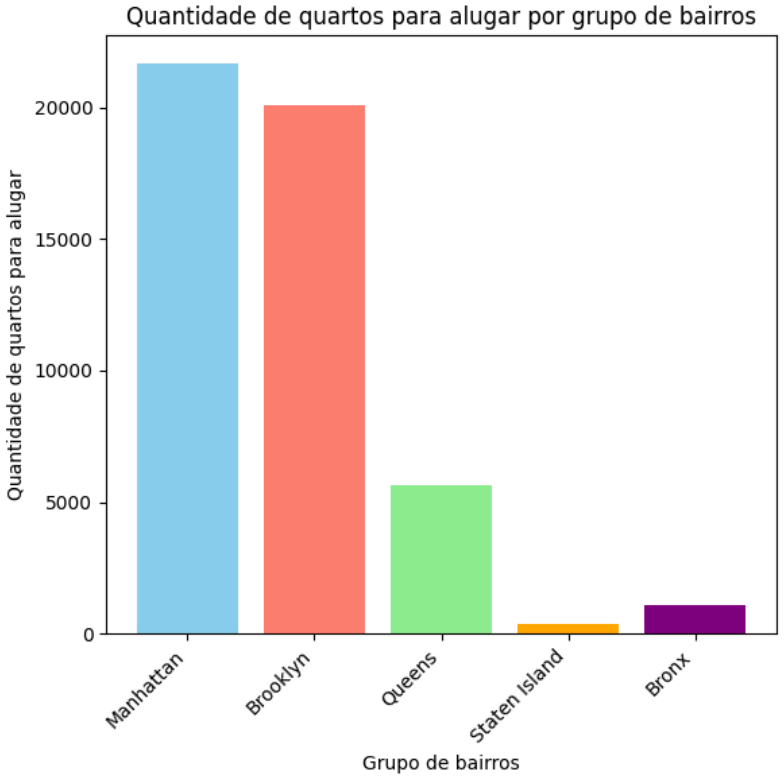
Para começar nossa análise centrada na variável de preço, é crucial examinar sua distribuição. Isso nos dará insights valiosos sobre a faixa de preços dos aluguéis e sua dispersão. Vamos prosseguir com a análise da distribuição da variável de preço.



Ao analisar o boxplot da distribuição de preços, destaca-se a presença de valores atípicos nesta base de dados.

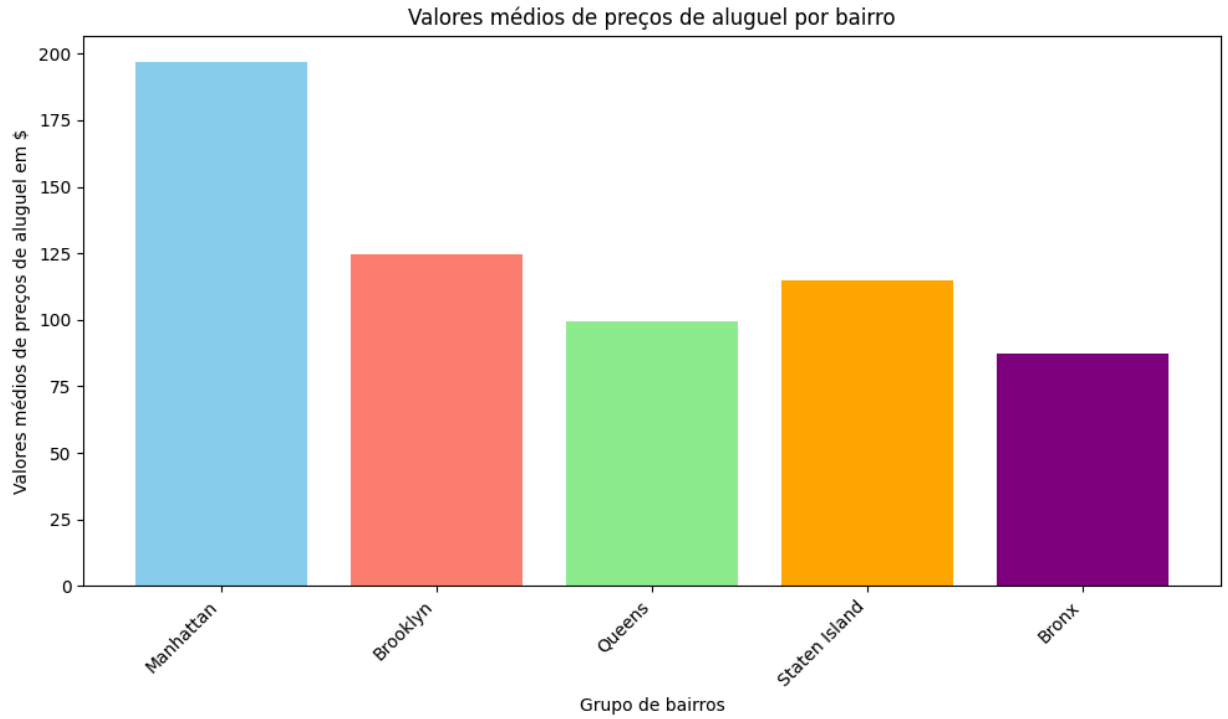


Em uma segunda análise, após remover esses valores atípicos, podemos obter uma visão mais precisa da distribuição dos preços, concentrando-nos na faixa de 60 a 180. Isso nos permitirá examinar com mais detalhes a distribuição dos preços dentro dessa faixa mais relevante para nossa análise.



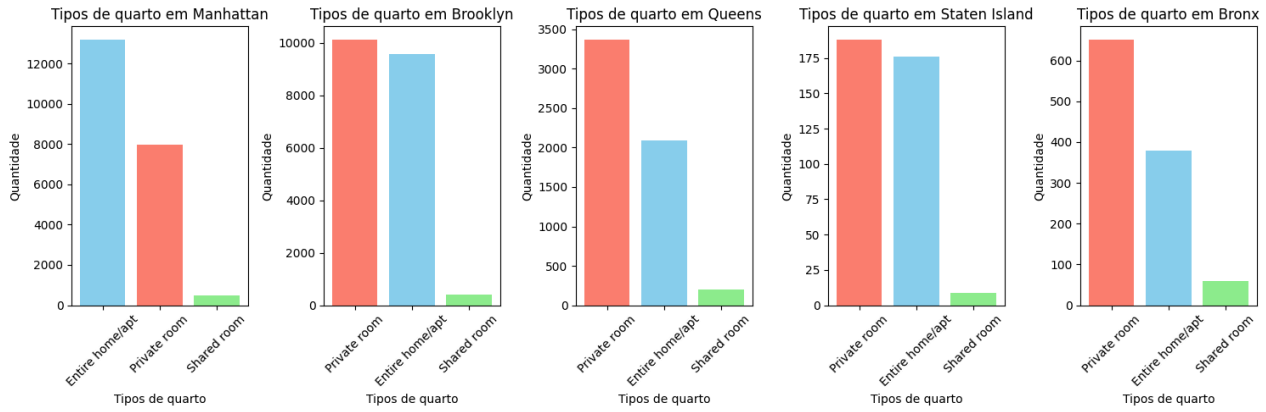
A distribuição de quartos para alugar é a seguinte:

21.642 em Manhattan
20.079 no Brooklyn
5.664 no Queens
373 em Staten Island
1.088 no Bronx



A distribuição de preços de quartos para alugar é a seguinte:

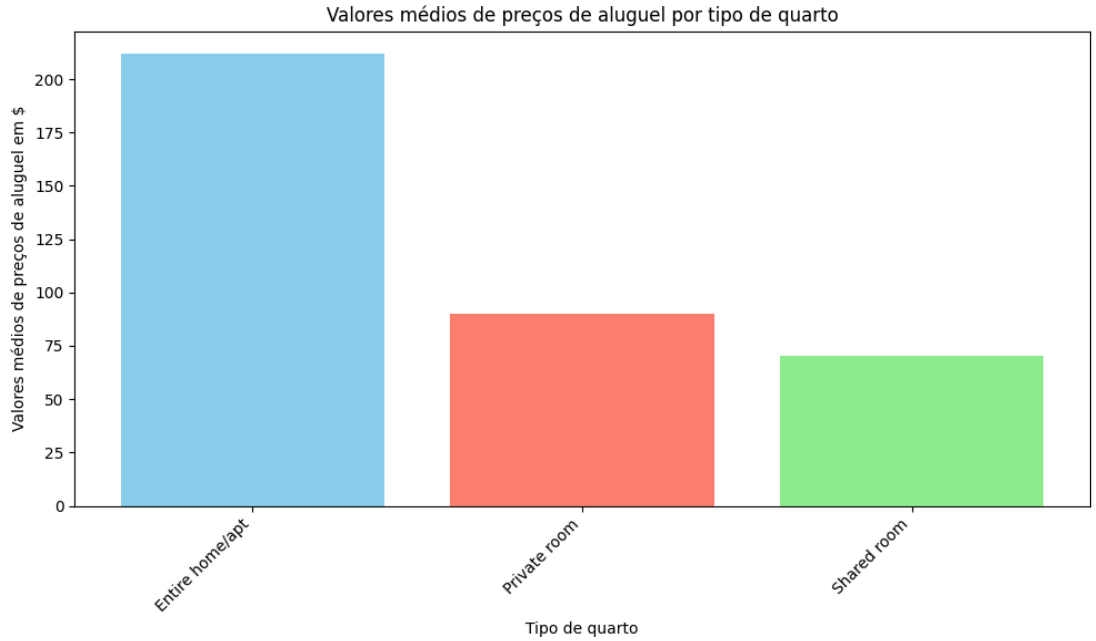
$196.90 em Manhattan
$124.46 no Brooklyn
$99.53 no Queens
$114.81 em Staten Island
$87.54 no Bronx



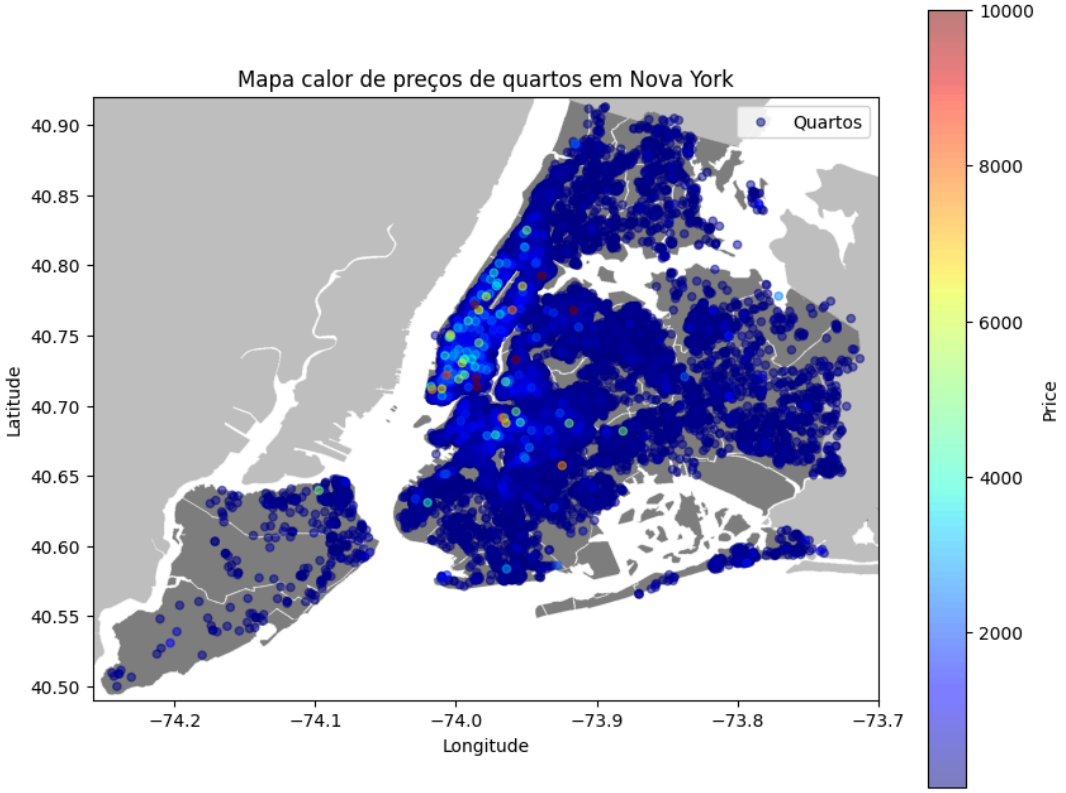
Analisando a distribuição dos tipos de quartos, podemos observar que Manhattan é o único grupo de bairro em que há mais "Entire home/apt" do que "Private room". Os números absolutos são os seguintes:

Tabela 4 - Quantidade de cada tipo de quarto nos grupos de bairros

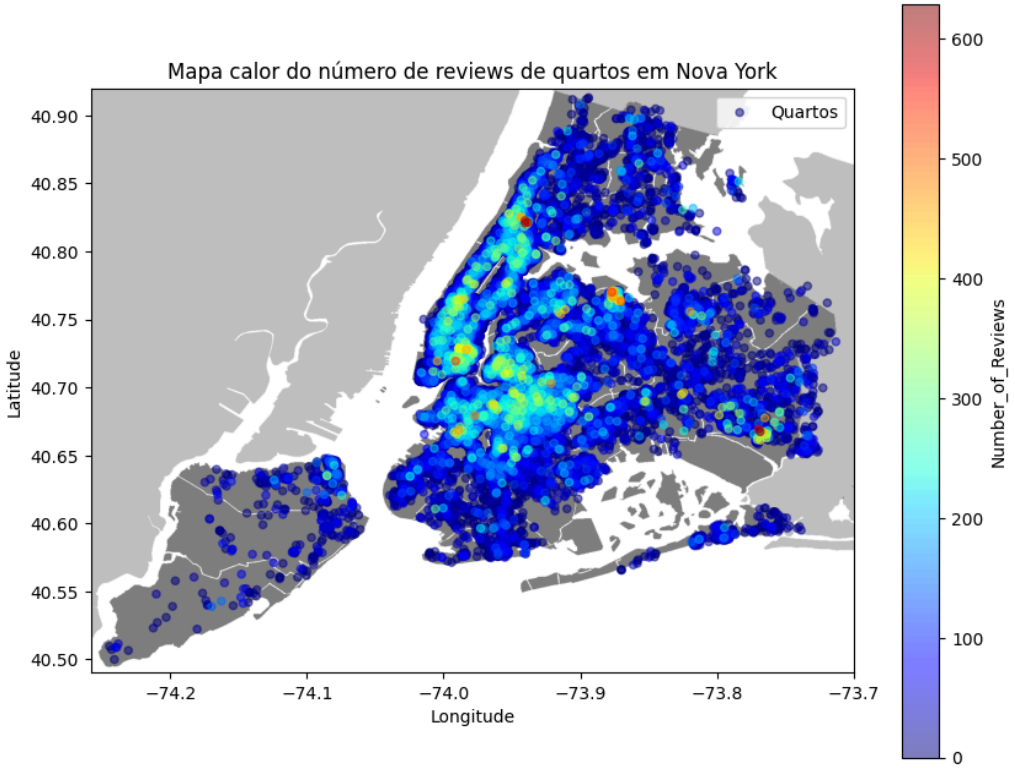
	Manhattan	Brooklyn	Queens	Staten Island	Bronx
Entire home/apt	13189	9552	2096	176	378
Private room	7973	10116	3370	188	651
Shared home	480	411	198	9	59



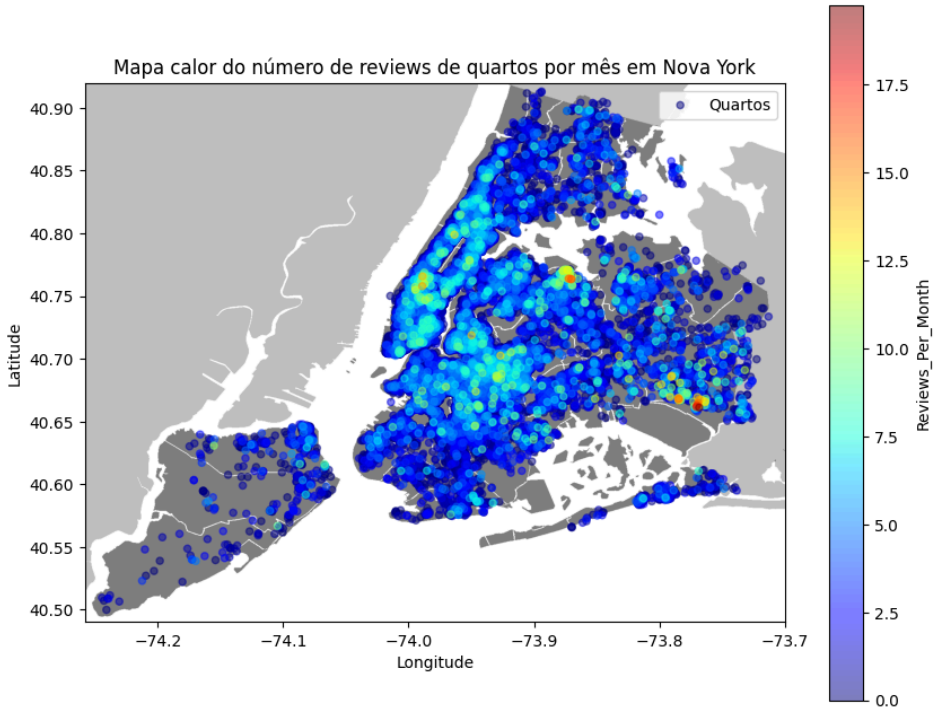
Ao analisar os valores médios para cada tipo de quarto, observa-se uma disparidade significativa entre os "Entire home/apt" e os outros tipos. Os valores médios são os seguintes: $211.82 para "Entire home/apt", $89.81 para "Private room" e $70.19 para "Shared room".



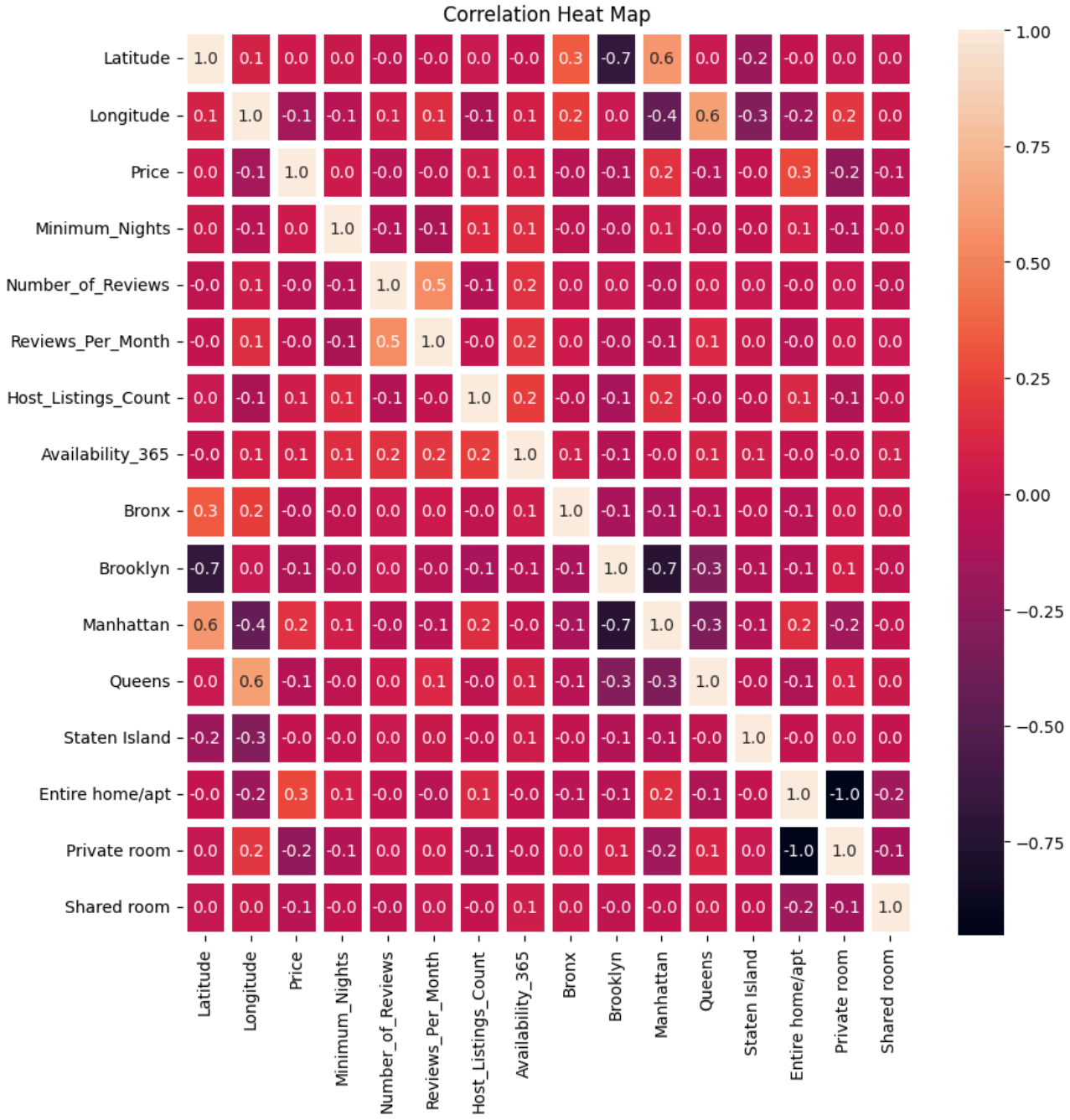
Analisando de forma mais visual com um mapa de Nova York, é evidente que Manhattan é o distrito com os valores mais altos de aluguel.



Ao analisar os números e avaliações dos quartos de aluguel, podemos identificar as áreas mais movimentadas e bem-sucedidas em termos de aluguel de quartos. Isso nos fornecerá insights sobre as partes da cidade que têm uma alta demanda por acomodações e que estão conseguindo efetivamente alugar esses quartos.



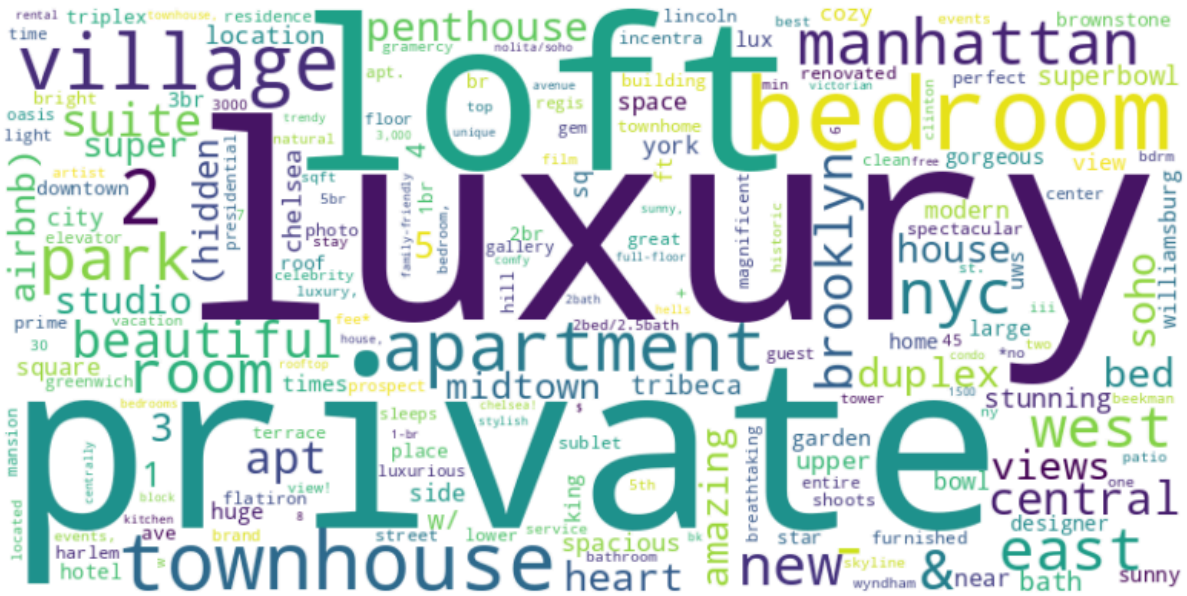
Ao analisar o número de avaliações por mês, podemos identificar os locais onde os quartos recebem avaliações de forma recorrente, indicando uma demanda consistente e possivelmente uma alta taxa de ocupação. Isso nos permite entender melhor as áreas onde os aluguéis são populares e têm um bom desempenho ao longo do tempo.



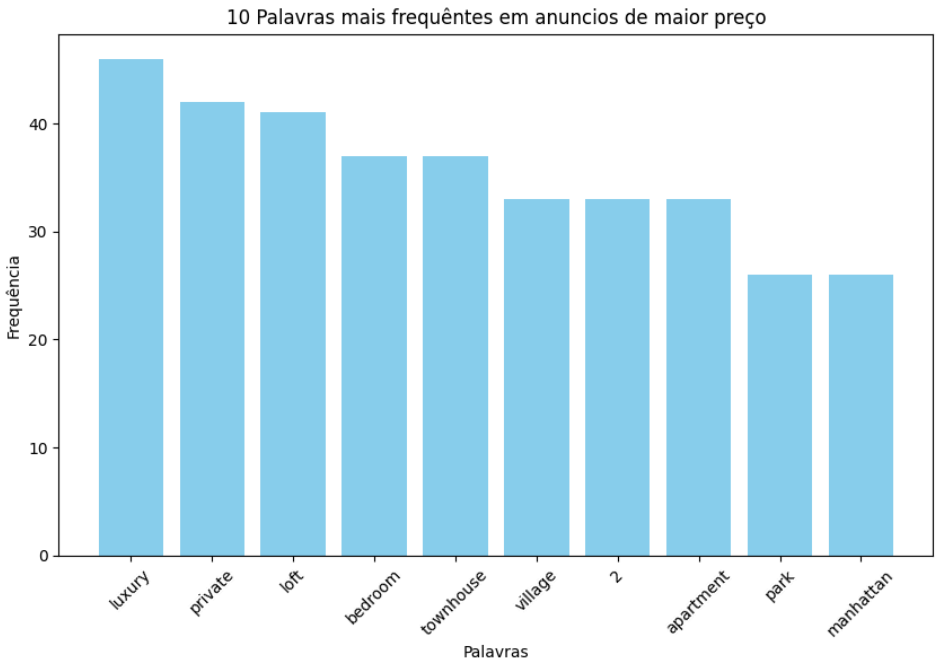
Ao analisar as correlações, observamos que o preço está correlacionado de forma relevante com poucas colunas. Isso sugere que apenas alguns fatores têm um impacto significativo no preço dos aluguéis, enquanto outros podem ter uma influência mínima ou nula.

Análise de Texto:

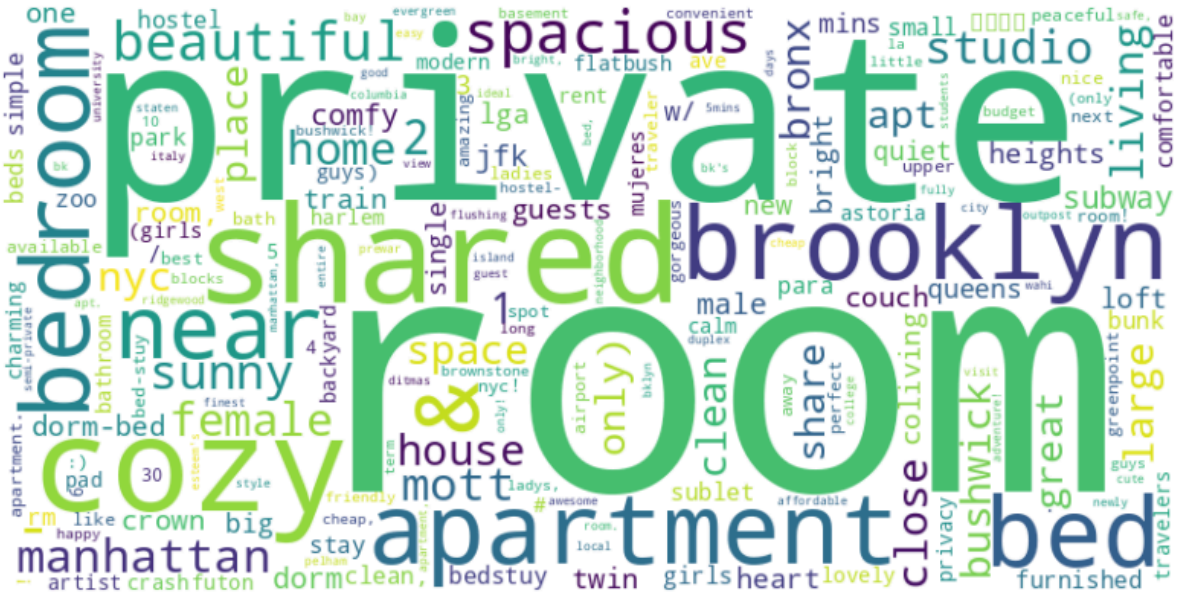
Essas são as palavras mais comuns nos anúncios



Essas são as palavras mais comuns em anúncios de alto preço acima de 800$ por dia.



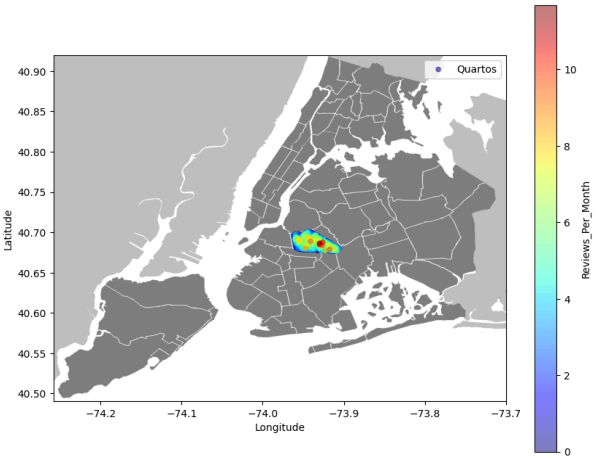
As palavras mais frequentes nos quartos com valores mais elevados incluem termos como "luxury", "private", "loft", "bedroom", "townhouse", "village", "apartment", "park" e "Manhattan". Essas palavras sugerem uma associação com características e localizações premium, o que pode influenciar os preços mais altos desses quartos.



Enquanto nos anúncios com aluguel de menor valor com o preço menor que $30, temos palavras diferentes, tais como ‘shared’ e ‘booklyn’.

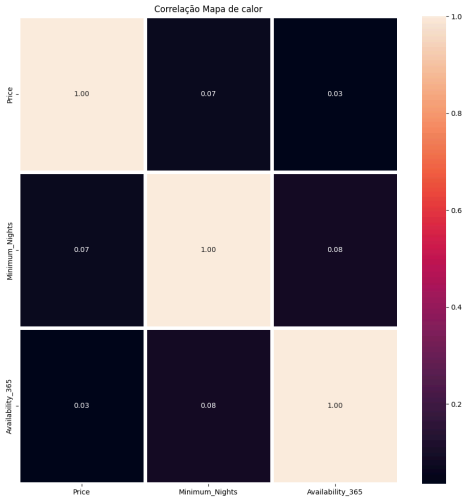
Conclusões

- Onde seria mais indicada a compra uma casa para alugar?



O bairro de East Elmhurst parece ser uma boa opção, uma vez que possui o maior número de casas com 10 avaliações por mês, o que sugere uma alta demanda na região. Além disso, a média de preço de aluguel no bairro é de $81.18. Isso indica que é uma área popular entre os locatários e pode oferecer oportunidades interessantes de investimento em aluguéis.

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



Com base na análise da matriz de correlação, observamos algumas relações muito fracas entre o preço, o número mínimo de noites e a disponibilidade. A correlação entre preço e número mínimo de noites é positiva, aproximadamente 0.066, o que sugere uma leve tendência para propriedades com preços mais altos terem

requisitos de estadia mínima ligeiramente mais longos. No entanto, essa correlação é bastante fraca.

Da mesma forma, a correlação entre preço e disponibilidade ao longo do ano é positiva, mas muito fraca, cerca de 0.035, indicando uma leve tendência para propriedades com preços mais altos terem um pouco mais de disponibilidade ao longo do ano. Mais uma vez, essa correlação é bastante fraca.

Em resumo, embora haja algumas associações leves entre preço, número mínimo de noites e disponibilidade, nenhuma delas é forte o suficiente para indicar uma relação significativa entre essas variáveis.

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

Definitivamente, há um padrão perceptível nos nomes dos locais com valores mais altos. Palavras como 'luxury', que sugerem lugares de maior valor, e 'Manhattan', que está associada ao grupo de bairros com a maior média de preço, são pontos importantes a serem destacados. Além disso, a presença frequente do número 2 nas palavras mais comuns pode indicar que esses locais geralmente têm uma quantidade maior de espaços ou outras características adicionais, o que pode influenciar os preços mais altos. Esses insights são valiosos para entender as características e os fatores que contribuem para os valores mais elevados nos aluguéis.

- Quais variáveis mais tem relação com o preço?

As variáveis com maior relação com preço a princípio são de grupo de bairro e de tipo de quarto, apesar de estas ainda não terem uma relação direta tão forte comparado com as outras são as mais forte.

- O tipo de quarto influencia diretamente o preço?
  Sim, ao fazer análise nota-se que a média de preço dos quartos tem uma grande diferença

- O distrito com mais quartos de alto valor é o distrito com a maior média de preços.

Se comprovou que o único distrito em que a maioria dos quartos é o que tem a maior média de preços é o distrito com maior média de preços.

- Os quartos de menor valor tem algum padrão de texto

Sim, se comprovou que assim como os quartos de maior valor os de menor valor também seguem um padrão de nome de anúncio o que torna possível em um eventual desenvolvimento de um modelo preditor, o uso de tokenização de palavras para predizer.