Desafio Lighthouse

Objetivo:

O Objetivo desse desafio é identificar e desenvolver uma estratégia de precificação competitiva, através de uma análise exploratória dos dados de uma plataforma de aluguéis temporários na cidade de Nova York, além do desenvolvimento de um modelo preditivo de preços.

Dicionário de Dados:

id – Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo
nome - Representa o nome do anúncio
host_id - Representa o id do usuário que hospedou o anúncio
host_name – Contém o nome do usuário que hospedou o anúncio
bairro_group - Contém o nome do bairro onde o anúncio está localizado
bairro - Contém o nome da área onde o anúncio está localizado
latitude - Contém a latitude do local
longitude - Contém a longitude do local
room_type – Contém o tipo de espaço de cada anúncio
price - Contém o preço por noite em dólares listado pelo anfitrião
minimo_noites - Contém o número mínimo de noites que o usuário deve reservar
numero_de_reviews - Contém o número de comentários dados a cada listagem
ultima_review - Contém a data da última revisão dada à listagem
reviews_por_mes - Contém o número de avaliações fornecidas por mês
calculado_host_listings_count - Contém a quantidade de listagem por host
disponibilidade_365 - Contém o número de dias em que o anúncio está disponível para reserva

Análise Exploratória dos Dados (EDA)

1.1 Tratamento

Antes de realizar a análise é necessário fazer os tratamentos dos dados, nisso inclui identificar se há valores vazios, e conceber uma maneira de tratar de esses valores, caso ocorram, de modo que não cause prejuízos a interpretação, como também conclusões errôneas ou falhas.

No tocante a isso, foram identificados valores vazios em várias colunas no data-set, sendo elas as colunas de “nome”, “host_name”, “ultima_review” e “reviews por mês”. Abaixo segue uma tabela, onde é possível identificar essas ausências.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48894 entries, 0 to 48893
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48894 non-null  int64  
 1   nome                            48878 non-null  object 
 2   host_id                         48894 non-null  int64  
 3   host_name                       48873 non-null  object 
 4   bairro_group                    48894 non-null  object 
 5   bairro                          48894 non-null  object 
 6   latitude                        48894 non-null  float64
 7   longitude                       48894 non-null  float64
 8   room_type                       48894 non-null  object 
 9   price                           48894 non-null  int64  
 10  minimo_noites                   48894 non-null  int64  
 11  numero_de_reviews               48894 non-null  int64  
 12  ultima_review                   38842 non-null  object 
 13  reviews_por_mes                 38842 non-null  float64
 14  calculado_host_listings_count  48894 non-null  int64  
 15  disponibilidade_365             48894 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
None
```

Cada um dos casos em que esse problema foi identificado teve maneiras distintas de tratar. Para os valores ausentes na coluna de nome, optou-se por remover essas entradas, uma vez que elas representam menos de 1% do banco de dados. Enquanto para “host_name”, foi feito apenas uma troca por “unknown”.

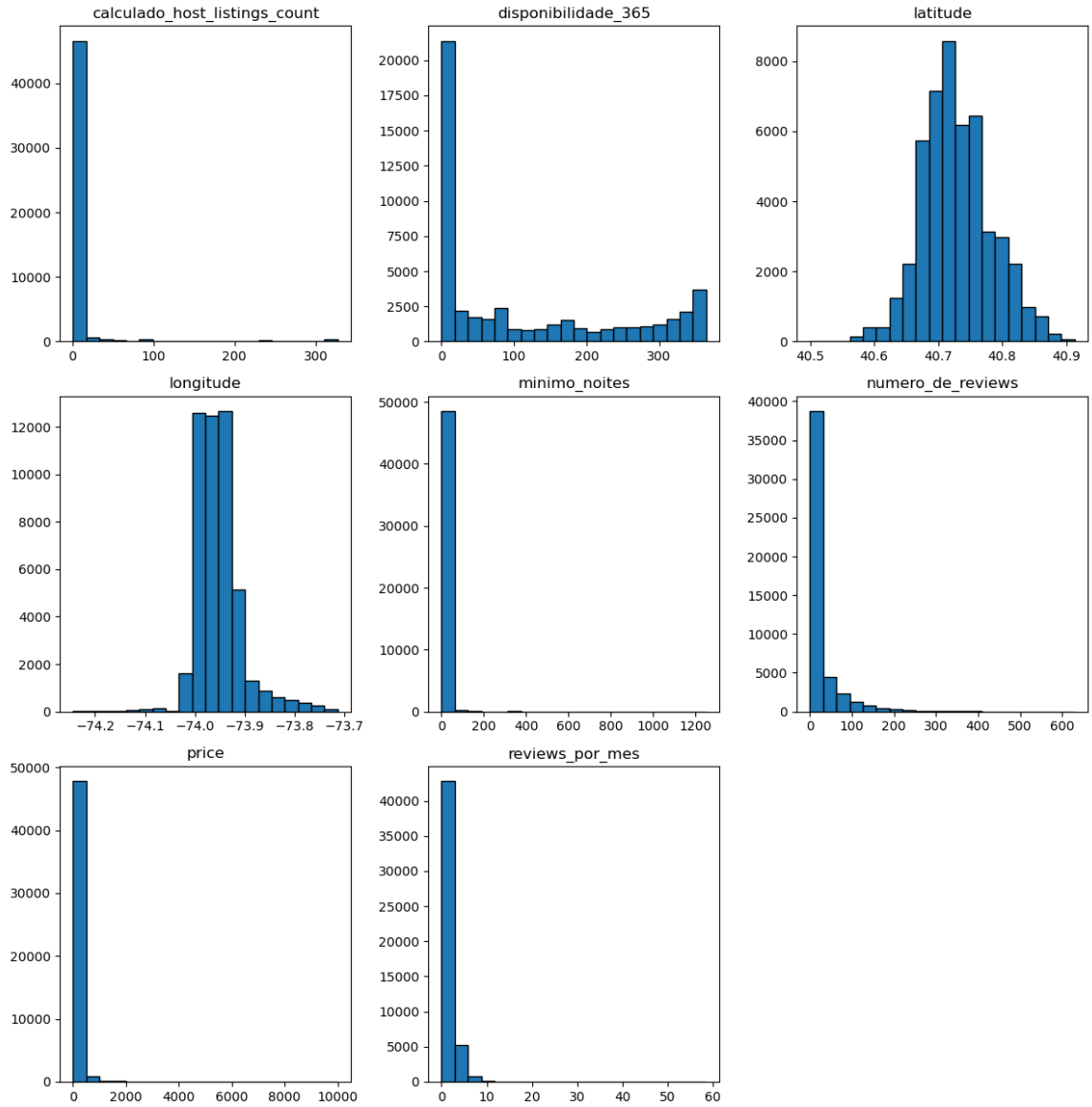
Na coluna de “ultima_review”, optou-se por substituir o campo vazio pela data mais antiga no bloco, pois essa situação poderia ter sido causada devido a necessidade de atualização do anuncio, como também poderia ser provocado pelas dificuldades de locação. O mesmo problema foi identificado em “reviews_por_mes”, no mesmo volume, para essa situação, fez-se a substituição por zero.

Outra anomalia foi identificada no data-set, alguns anúncios apresentavam preço igual a zero. Para realizar o tratamento, calculou-se a média dos preços da região mais próxima, dentro dos bairros (bairro) e da região maior (bairro_group). Então

seria feita a substituição, primeiramente pelas regiões mais próximas, e, caso o valor continue-se sendo nulo, pelo preço médio das regiões macro.

1.2 Identificando Outliers

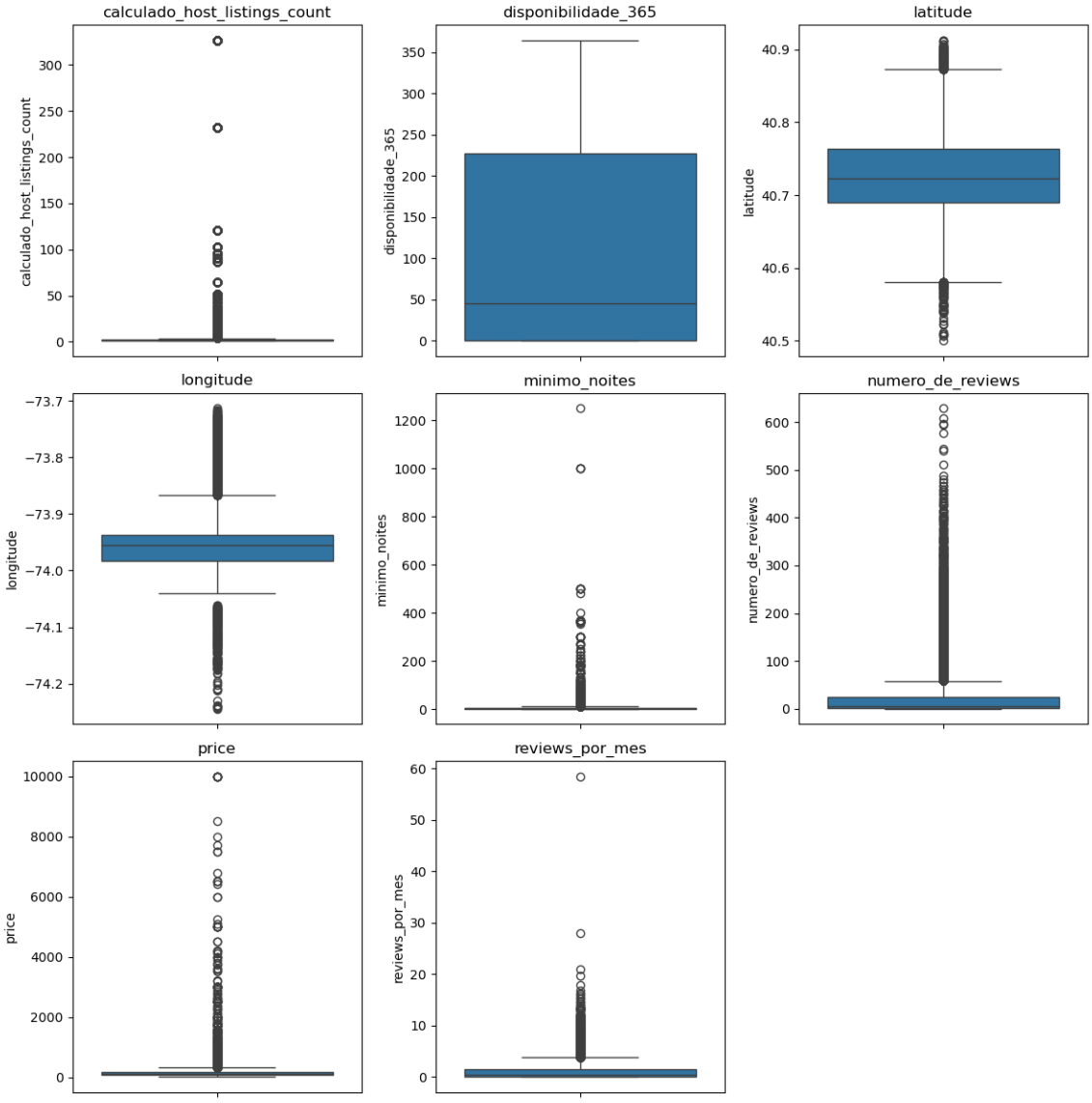
Feitos os devidos tratamentos, prosseguiu-se na análise, buscando compreender mais a distribuição das variáveis numéricas. Com esse objetivo, foi realizado o estudo de histogramas dessas variáveis.



Os histogramas apresentados evidenciam a grande heterogeneidade das variáveis analisadas, uma vez que a maioria dos gráficos apresenta assimetria e se distancia de uma distribuição normal. Essa característica era esperada, considerando o escopo dos dados, a complexidade do cenário socioeconômico da cidade de Nova York, as particularidades de cada região e a alta variabilidade de preços nesse mercado.

Os gráficos de boxplot refletem esse mesmo padrão, porém oferecem uma compreensão mais detalhada sobre os valores extremos em cada variável. Observa-se uma grande variação nos preços, bem como a presença de outliers, tanto no preço quanto no número mínimo de noites (mínimo_noites).

Esses valores discrepantes podem influenciar a escolha e o desempenho do futuro modelo de predição a ser desenvolvido. No entanto, para evitar impactos na análise exploratória dos dados ou distorções nas interpretações deste estudo, optou-se por não realizar um tratamento específico para outliers neste momento.

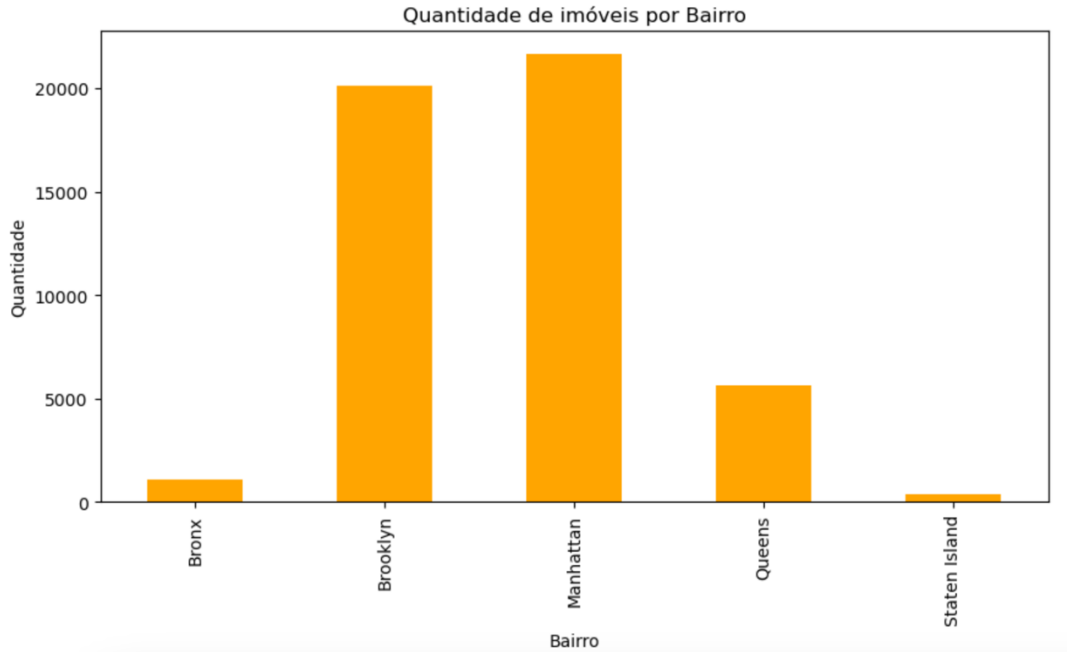


Após essa constatação, tem-se o prosseguimento do estudo, dando-se início a uma análise mais detalhada sobre cada variável.

1.3 Análise das Variáveis

1.3.1 Bairros Principais (bairro_group)

1.3.1.1 Quantidade de Imóveis



É possível perceber, por meio do gráfico, a distribuição dos anúncios por bairro. Manhattan é o bairro com a maior quantidade de imóveis para locação, seguido pelo Brooklyn. Juntos, esses dois bairros representam a maioria dos imóveis anunciados na plataforma.

1.3.1.2 Acomodação mais comum entre os bairros

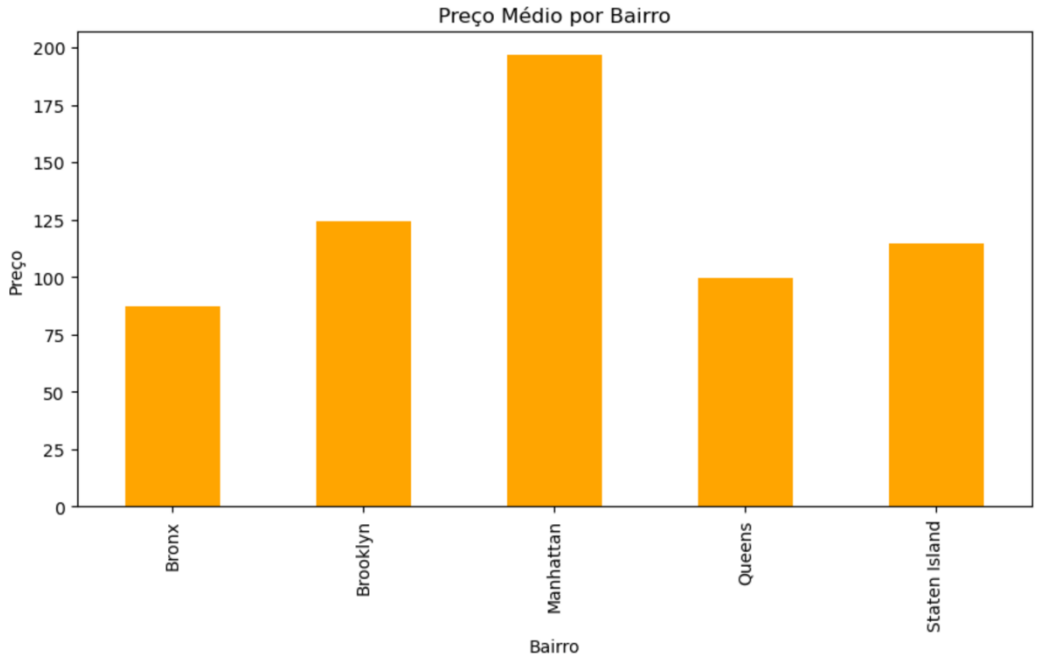
Pelos dados da tabela, é perceptível a predominância quartos privativos na maioria dos bairros. A única exceção é Manhattan, tendo predomínio por Casa/Apartamento inteiro.

	Bairro	Tipo Mais Comum	Quantidade
1	Manhattan	Entire home/apt	13193
2	Brooklyn	Private room	10126
3	Queens	Private room	3372
4	Staten Island	Private room	188
5	Bronx	Private room	652

1.3.1.3 Média de preços por bairro.

O gráfico permite visualizar como os preços médios de imóveis variam entre os diferentes bairros da cidade. Manhattan apresenta o maior preço médio entre

todos os bairros. Essa alta pode ser explicada por diversos fatores, como localização privilegiada, maior oferta de serviços e infraestrutura, além de uma demanda mais elevada por imóveis. Brooklyn e Staten Island apresentam preços médios similares, isto pode indicar que esses bairros apresentam muitas semelhanças entre si. Bronx e Queens são os bairros com o menor preço médio, sendo Bronx o com menor.

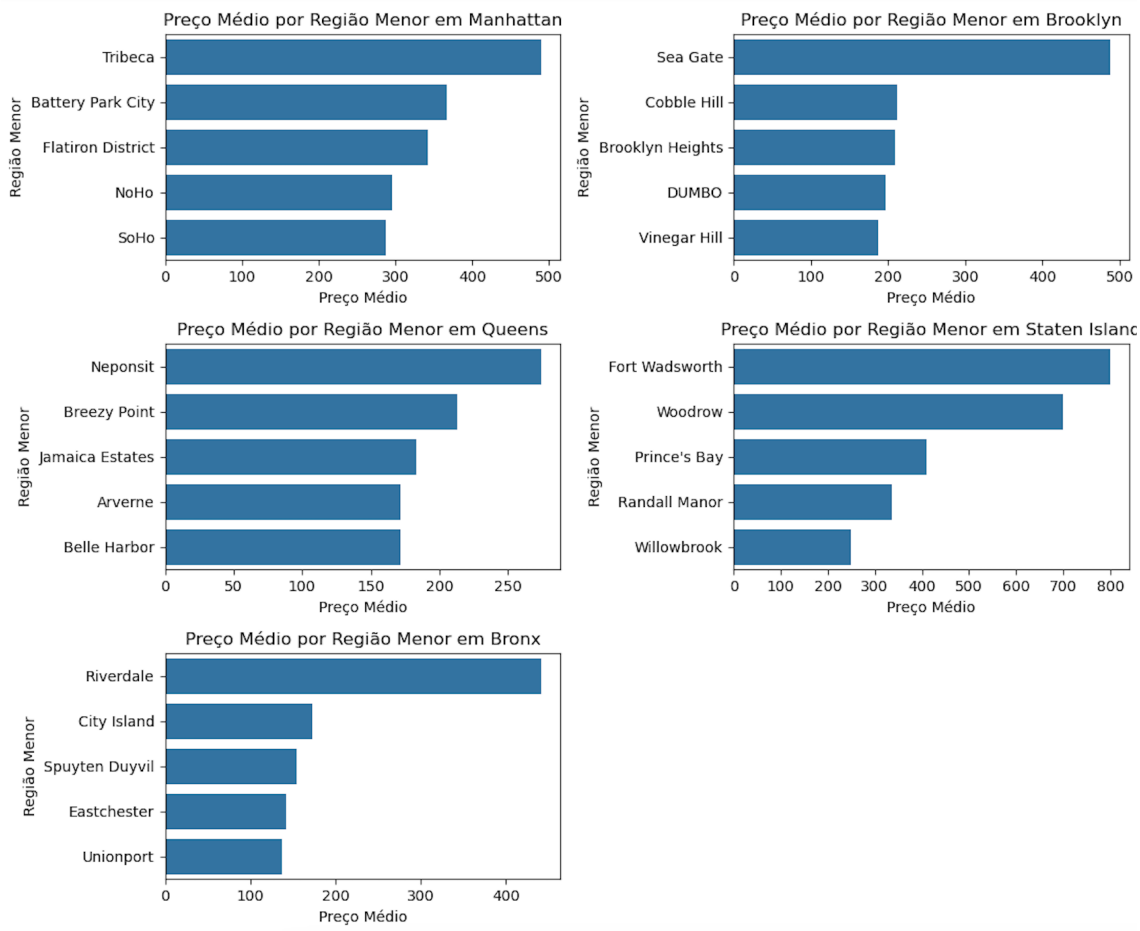


1.4 Regiões Menores(bairros)

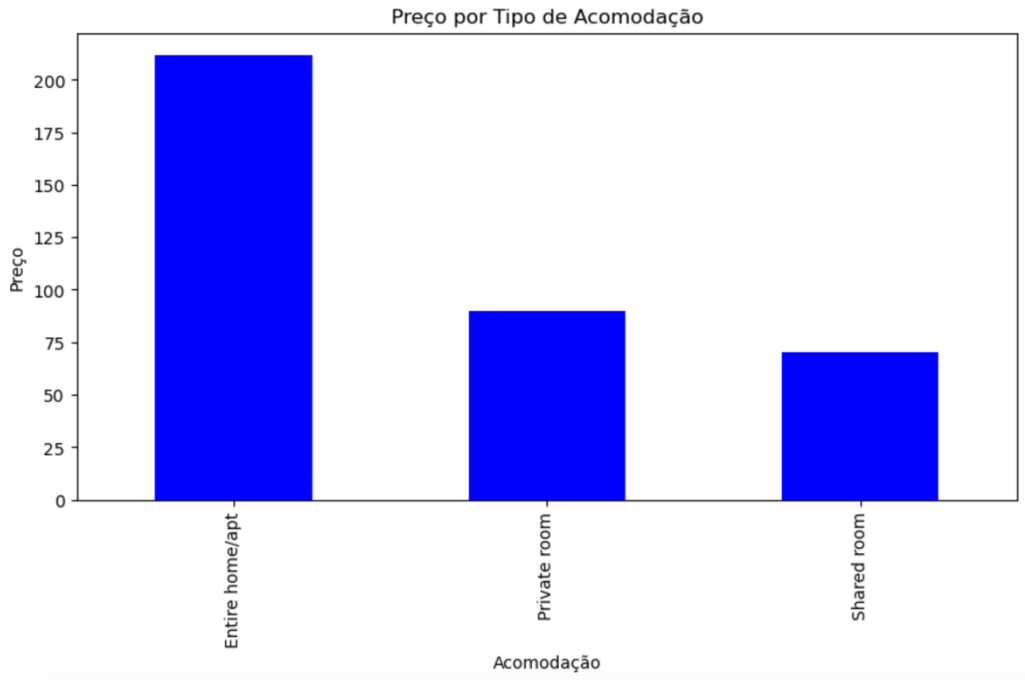
Observando os gráficos abaixo, é notável variação dos preços médios dentro das diferentes regiões menores dentro de cada bairro, ainda que a análise compreenda as 5 regiões menores, de preço médio mais elevado. Isso indica que a localização dentro do bairro é um fator determinante no valor dos imóveis, além de evidenciar que a heterogeneidade é uma característica comum do mercado imobiliário de Nova York.

Essa variabilidade pode causar prejuízos a análises nas quais se considere apenas o preço médio por bairro. No entanto, deve-se ressaltar, Manhattan aparenta ter uma das menores variabilidade dentro os bairros, mesmo contento regiões extremamente valorizadas, o que explica essa área ter o preço médio mais valorizado.

Staten Island a região com preço médio mais valorizada dentre todo o grupo. Porém, seu preço médio é inferior a Manhattan. A grande variação não só explica isso, como também fica demontrada pelo gráfico. Os preços médios em Bronx são os mais baixos entre os cinco bairros, com uma menor variação entre as diferentes regiões menores.



1.5 Tipo da Acomodação



A acomodação mais anuncia com maior valor é “Casa/Apartamento inteiro”. Esse preço pode ser explicado por alguns elementos como, exclusividade de ter uma casa inteiro à disposição, com todas as comodidades e privacidade. O valor agregado por esse tipo de acomodação, representa mais de 50% em relação aos anúncios para outros tipos de locação temporária. O menor valor está no quarto compartilhado, isso é facilmente explicado pela dificuldade de privacidade que as pessoas têm nesses lugares.

1.6 Número Mínimo de Noites

```
Minimo de Noites Registrado é: 1
Maior Minimo de Noites Registrado é: 1250
Média Minimo de Noites é: 7.01
count    48878.00
mean         7.01
std         20.02
min          1.00
25%          1.00
50%          3.00
75%          5.00
max       1250.00
Name: minimo_noites, dtype: float64
```

A maioria do Mínimo de noites está dentro de até 5 dias. O desvio padrão de 20, 02% é elevado, indicando a grande variabilidade dos dados. Isso é evidenciado pelo outlier de 1250 dias, o que afeta os resultados da média.

1.7 Número de Reviews

```
Número Máximo de Reviews Registrado: 629
Número Mínimo de Reviews Registrado: 0
Média de Reviews: 23.28
Quantidade de Reviews Mais frequente: 0
count    48878.00
mean        23.28
std         44.56
min          0.00
25%          1.00
50%          5.00
75%         24.00
max        629.00
Name: numero_de_reviews, dtype: float64
```

A grande maioria dos anúncios tem um número de reviews vai até 24. O desvio padrão é de 44,56%, igualmente elevado. Isso é provocado grande da variabilidade de dados. O valor máximo identificado é de 629 reviews, um outlier o qual afeta a interpretação de média dessa variável.

1.8 Data das Reviews

Data da Review Mais Recente: 2019-07-08
Data da Review Mais Antiga: 2011-03-28

1.9 Reviews por Mês

```
Maior Número de Reviews Registrados num Mês: 58.5
Média Registrada de Reviews num Mês: 1.09
Quantidade de Reviews Mais Frequente por Mês: 0.0
count    48878.00
mean         1.09
std          1.60
min          0.00
25%          0.04
50%          0.37
75%          1.58
max         58.50
Name: reviews_por_mes, dtype: float64
```

O número de reviews por mês varia consideravelmente, indo de 0 a 58,5. O desvio padrão é 1,60%, um valor alto, confirmando a grande variabilidade observada. A maioria dos anúncios apresentam um número de 1,58 reviews por mês. A presença de outliers influencia significativamente a média, sendo a mediana uma medida mais adequada para representar o centro da distribuição nesse caso.

1.10 Disponibilidades de Dias

```
Maior disponíbilidade: 365
Menor disponíbilidade: 0
Disponibilidade Mais Frequente: 0
count    48878.00
mean       112.78
std        131.61
min          0.00
25%          0.00
50%         45.00
75%        227.00
max        365.00
Name: disponibilidade_365, dtype: float64
```

A maioria dos anúncios para aluguel apresenta pouca ou nenhuma disponibilidade, indo desde 0 a 365. No primeiro quartil, as acomodações têm nenhuma disponibilidade. No Intervalo Interquartil, o cenário muda e passa-se a ter 45 de disponibilidade. E no terceiro quartil, 227 dias disponíveis. O desvio padrão é de 131,61, uma evidência da heterogeneidade dos dados. A presença de outliers afeta a média para cima.

1.11 Quantidade de listagem por host

```
Maior número de listagens: 327
Menor número de listagens: 1
Média de listagens: 7.15
count    48878.00
mean         7.15
std         32.96
min          1.00
25%          1.00
50%          1.00
75%          2.00
max        327.00
Name: calculado_host_listings_count, dtype: float64
```

A listagem de anúncios varia de 1 a 327, sendo que até 75% desses imóveis tem até 2 listagens, quando esse indicativo possui uma média de 7,15. Isso indica o impacto dos outliers nesse dado. O desvio padrão de 32,96 é uma evidência da variabilidade dos dados.

1.12 Preço

```
Maior Preço: 10000.0
Menor Preço: 10.0
Preço Médio: 152.75
Preço Mais Frequente: 100.0
count    48878.00
mean       152.75
std        240.18
min         10.00
25%         69.00
50%        106.00
75%        175.00
max      10000.00
Name: price, dtype: float64
```

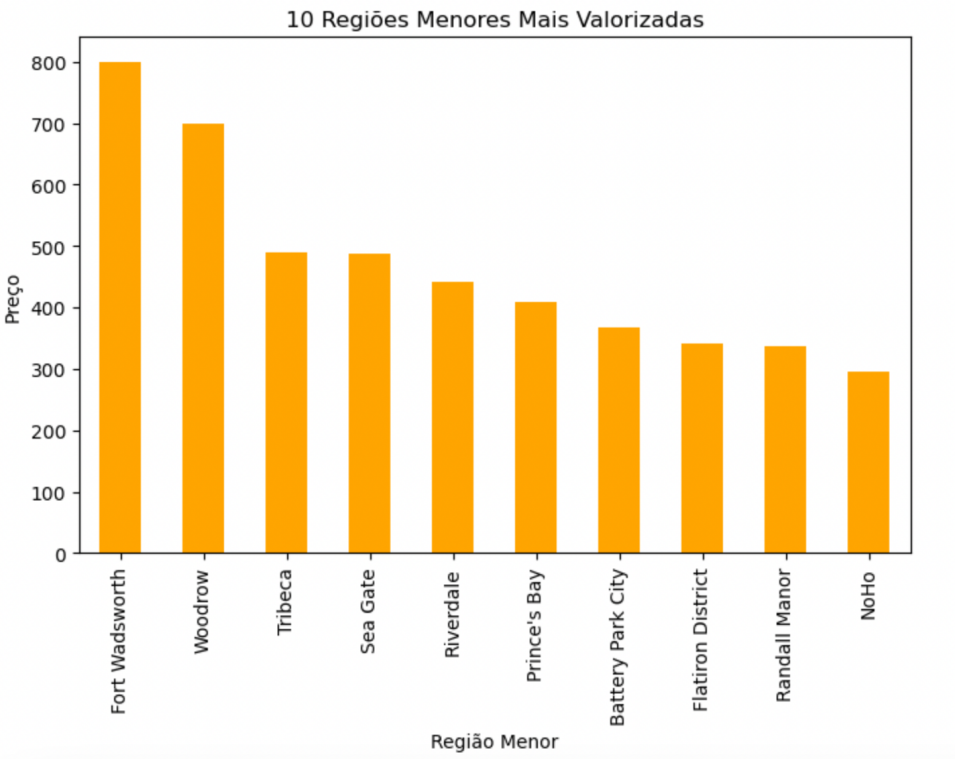
Os preços para o aluguel das acomodações variam desde 10 a 10000, isso evidencia a grande variabilidade dos dados, representado pelo elevado desvio padrão de 240,18. A maioria dos anúncios tem preço menor ou igual 175, porém a média para essa variável é de 152.75, esse valor é um impacto da presença de outliers nos dados.

1.13 Hipóteses:

- Tanto bairros, como também o tipo da acomodação, parecem serem fatores que impactam no preço do aluguel. Investir em imóveis do tipo “Casa/Apartamento inteiro”, localizados em bairros com aluguéis mais caros, pode resultar em retornos maiores.
- O número de listagens pode impactar no preço, acomodações com mais listagens parecem ter menos disponibilidade. Devido a grande procura, o anfitrião pode aumentar o preço do aluguel.

- Ter um número mínimo de noites elevado pode impactar no preço, um período maior, faz com seja necessário alugar a acomodação por mais tempo, o que pode impactar no preço do aluguel.

2.1 Sugestão de compra:

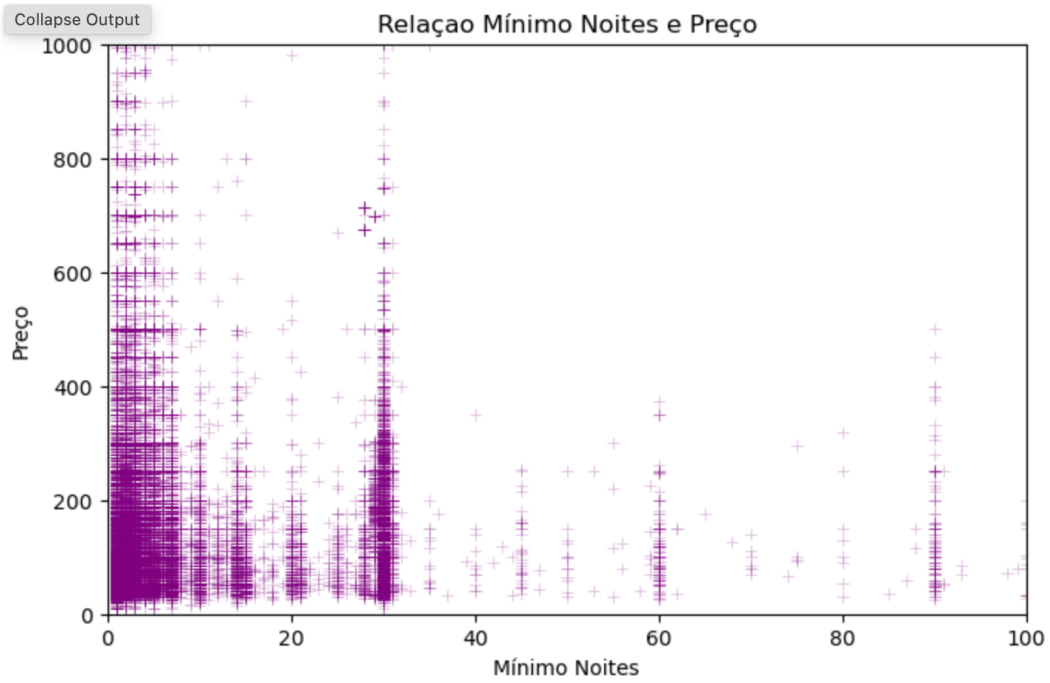


Para um investimento mais seguro, o mais indicado seria investir em um apartamento em Manhattan, devido a maior concentração de imóveis nesse bairro, e devido ao maior preço médio das entre todos os bairros.

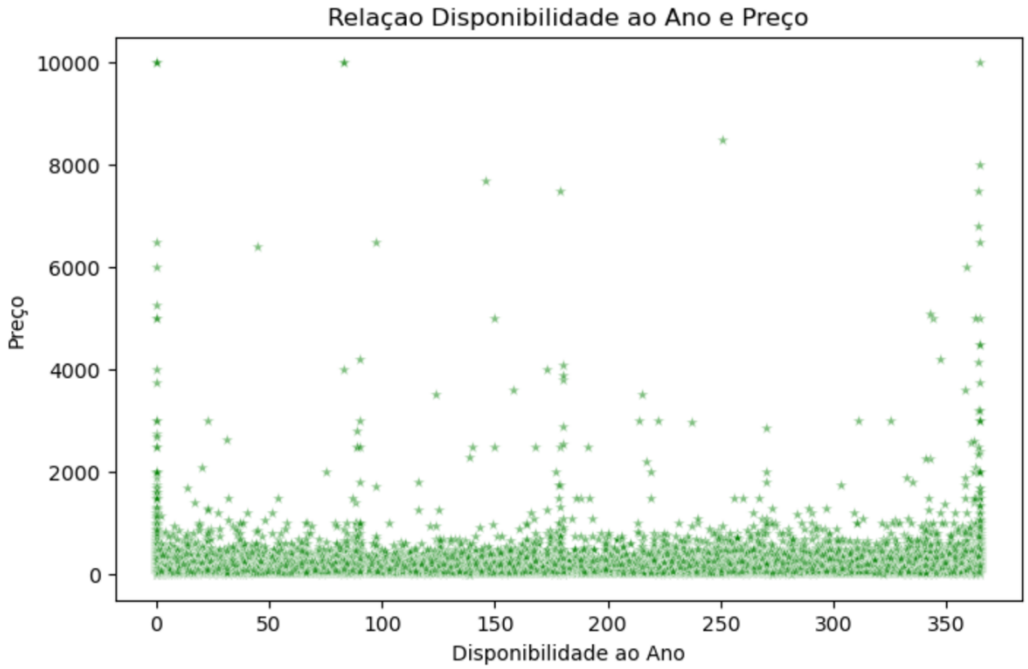
Contudo, se o investidor se está mais propenso a tomar riscos, talvez seja mais indicado ele investir no bairro de Staten Island, essa região tem 3 localidades entre as 10 mais valorizadas de Nova York, a primeira, Fort Wadsworth, e a segunda, Woodrow, ambas localizadas nesse bairro.

No entanto, esse bairro é o com menor concentração de imóveis, e possui um preço médio inferior a Brooklyn e Manhattan. A possibilidade de lucro é maior, porém o risco também.

2.2 Relação entre mínimo de noites e disponibilidade com preço:



Pelo gráfico, é possível observar que não existe uma relação clara entre Mínimo de Noites e Preço, uma vez que é possível constatar que conforme aumenta-se o mínimo de dias para o aluguel, o preço não apresenta um crescimento constante, e sim variabilidade, com picos no preço até mesmo menores.



Analisando o gráfico, é possível observar anúncios com imóveis que possuem nenhuma disponibilidade, ou seja, há grande procura, e imóveis os quais estão disponíveis durante todo, ambos aparentar estar relacionados a preços mais elevados. Contudo, exceto nas extremidades do eixo x, os dados do gráfico estão muito dispersos, indicado que existem outras variáveis mais relevante que impactam no preço.

2.3 Relação texto do nome do local e lugares com valor mais alto:

Com uma análise inicial dos dados, foi possível perceber uma relação de certas palavras chaves e o valor de tais localidades.

Foi possível perceber o efeito de nomes relacionados a edificações de alto valor agregado, tais como “Mansion”, “Penthouse”, “Townhouse”, sobre o preço. Cada um destes lugares, apresentou um preço médio superiores aos imóveis sem essas palavras chaves.

Um fator observado que também teve impacto, foi no nome do imóvel ter referência a grandes eventos, como o Superbowl, uma referência proximidade com o evento impacta no preço de forma positiva, valorizando o imóvel.

Outros termos chaves de impacto foram “Luxuary” e “Hidden hidden by airbnb”, ambos os termos também estão relacionados com acomodações com preços mais elevados.

3. Modelo

3.1 Correção das Variáveis

O mapa de calor apresentado mostra as relações dentre variáveis com preço. As cores quentes indicam uma correlação positiva forte, logo, quando uma variável aumenta, a outra também tende a aumentar. Enquanto as cores frias, indicam uma correlação negativa forte, logo, quando uma variável aumenta, outra tende a cair. Saber a correlação entre cada variável, pode auxiliar no desenvolvimento do modelo.

Neste tocante, existe uma correlação positiva fraca entre preço e número de reviews, e, também, uma correlação positiva fraca entre disponibilidade e preço. Há uma ausência de correlações fortes, isso indica que as variáveis analisadas são influenciadas por diversos fatores.

3.2 Regressão ou Classificação

Prever o preço de um produto ou serviço é um problema de regressão, pois, o objetivo é determinar um valor numérico específico e contínuo para uma instância futura. Já na classificação, buscasse atribuir uma classe ou categoria a um dado, a regressão visa prever um valor exato dentro de um intervalo numérico. No caso da previsão de preços, o objetivo é encontrar o valor mais provável que um determinado produto ou serviço terá no futuro, considerando um conjunto de características e dados históricos.

3.3 Desenvolvimento do Modelo

Para o desenvolvimento do modelo foi testado dois métodos predição:

3.3.1 Random Forest

O primeiro foi RandomForest, no qual consiste em criar vária arvores de decisão de maneira aleatória, combinando a decisão de cada uma das arvores de modo que aumente a precisão de seus resultados. Esse modelo é especialmente adequado para lidar com dados outliers, o caso de nosso banco de dados. Ele também é interessante quando se tem dificuldade de mensura a correlação entre variáveis.

Resultados Primários:

```
Random Forest
104.29
mae: 65.82
mse: 43306.00
rmse: 208.10
r2: 0.05
```

3.3.2 Regressão Linear

O segundo modelo foi o de Regressão linear, uma técnica a qual se baseia na suposição de que há uma relação linear entre a variável dependente e as variáveis explicativas. Para a regressão linear é importante haver uniformidade dos dados. Nela buscar-se prever o valor de uma variável dependente através das variáveis explicativas, capturando tendências e padrões nos dados.

Resultados Primários:

```
Regressão Linear
129.5942666530973
mae: 69.96
mse: 39894.09
rmse: 199.74
r2: 0.12
```

3.3.3 Análise de Resultados

Os dois modelos apresentam muitos problemas, ambos possuem a estatística mae muita elevada, prejudicando nos resultados finais. O r-quadrado é igualmente baixo, o que significa que os dois modelos explicam pouco a variável preço. Porém, na regressão linear, esse valor é um pouco mais alto, sendo ele o modelo o qual será escolhido para a próxima previsão.

4. Resultados Finais

Após a utilização do modelo, o valor previsto foi de 323.02, esse valor esta acima para os valores comparativos desses dados, uma representação da falha modelo, devido a heterogeneidade dos dados, porém que ao não serem tratados permitiu uma análise mais abrangente.

5. Conclusão

Essa análise poderia aprimorada utilizando outros dados, como dados socioeconômicos, um levantamento sobre aspectos econômicos da cidade de Nova York quando recebe grandes eventos, entre outros elementos e afins. A complexidade da cidade foi refletida nesses, sendo necessário uma abrangência maior de variáveis para compreensão e previsão, sobre os elementos que afetam o preço do aluguel em Nova York.