Relatório de Análise Exploratória dos Dados

Desafio Lighthouse - Indicium

Emanuelle Ferraz Lima

1. Introdução

A Análise Exploratória dos Dados (EDA) é um dos métodos mais importantes ao construir um modelo de Machine Learning (ML). A etapa de EDA é feita para que o cientista de dados possa conhecer a base de dados, entender o contexto por trás dos dados, e analisar cada um dos atributos presentes em um determinado dataset, resumindo assim esse conjunto de dados.

Nessa perspectiva, durante a realização do desafio de ciência de dados (Lighthouse) da indicium a primeira entrega a ser feita era justamente uma análise exploratória dos dados do dataset fornecido pela empresa. O objetivo do desafio era realizar a precificação do preço de aluguéis de imóveis do estado de Nova Iorque, caracterizando-se como um problema de regressão.

Dessa forma, a principal técnica a ser feita para conhecer e entender as variáveis presentes é a etapa de EDA, na qual foi divida entre analisar os atributos numéricos, categóricos, correlações entre os atributos e por último a verificação da existência de dados nulos e faltantes, como será mostrado a seguir. Destarte, no presente relatório estão descritos os principais insights e características das variáveis obtidas durante a EDA.

2. Objetivo e Contexto

Antes de começar a mostrar a etapa de análise exploratória dos dados, é preciso deixar claro que o objetivo do desafio é desenvolver um modelo de machine learning que faça a previsão dos preços de aluguéis de imóveis no estado de Nova Iorque, utilizando regressão e posteriormente avaliando com as devidas métricas a performance do modelo. Contudo, o primeiro passo quando desenvolve-se um modelo de ML é conhecer a sua base de dados, e é sobre isso que se trata esse relatório.

2.1 Conhecendo as variáveis do dataset:

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

3. Estatística Descritiva

Para analisar estatisticamente as variáveis do dataset, utilizei o método *describe()* do pandas, na qual foram obtidos alguns números interessantes em algumas dessas variáveis, como:

- price:
    - Média de preço: $ 152.72
    - Valor máximo: $ 10000.00
    - Valor mínimo: $ 0
    - Desvio padrão: 240.15
    - Mediana: 106.00

Uma observação a ser feita é em relação a esse valor mínimo que teve algumas ocorrências no dataset. O valor **0** em um contexto de preço de aluguéis de imóveis não faz sentido, ou seja, esses registros foram removidos.

- minimo_noites:
    - Média: 7 noites
    - Valor máximo: 1250 noites
    - Valor mínimo: 1 noite
    - Desvio padrão: 20.51
    - Mediana: 3.00

- disponibilidade_365:
    - Média: ~112
    - Valor máximo: 365
    - Valor mínimo: 0
    - Desvio padrão: 131.61
    - Mediana: 45.00

- numero_de_reviews:
    - Média: 23.27
    - Valor máximo: 629.00
    - Valor mínimo: 0
    - Desvio padrão: 44.55
    - Mediana: 5.00

4. Analisando Variáveis Numéricas

Como dito anteriormente, EDA foi dividida entre variáveis numéricas e categóricas e inicialmente foi feita a análise nos atributos numéricos através de gráficos de *Boxplot, Histogramas, Gráficos de Dispersão e Correlação Linear e Monotônica*.

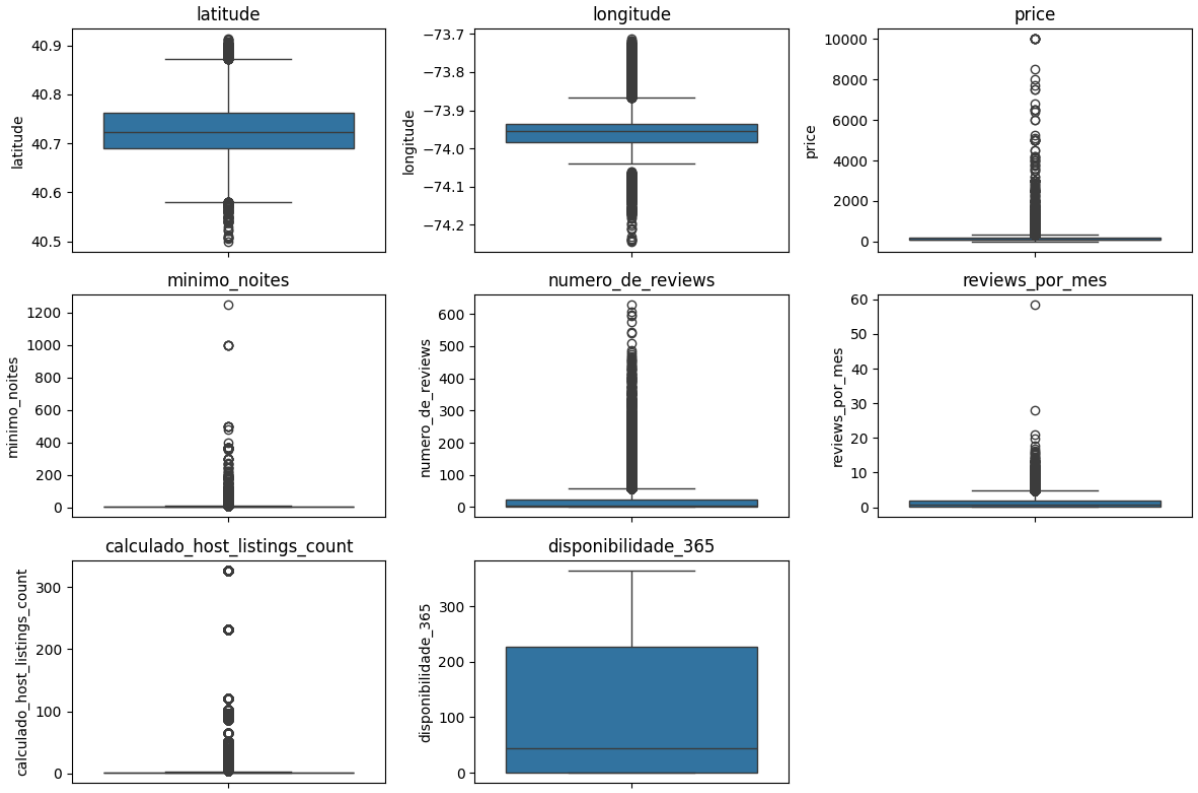


Figura 1: BoxPlot das variáveis numéricas

O gráfico de BoxPlot mostra visualmente a ocorrência de *outliers*, também conhecidos como anomalias, que são dados com valores muito maiores que o normal. Esse gráfico utiliza conceitos de estatística para definir limites inferiores e superiores para considerar registros como *outliers*, sendo eles: primeiro quartil, mediana e terceiro quartil, mínimo, máximo e o *IQR.* Importante ressaltar que o tratamento de *outliers* foi realizado apenas na variável price, retirando valores maiores que $9000 e observando se valores maiores que $6500 estavam corretos de acordo com as outras variáveis. Para verificar se os valores acima de $6500 estavam corretos, foram realizadas pesquisas sobre quais distritos de Nova Iorque tinham os preços mais altos e quais os bairros desses distritos costumam ter valores elevados também.

A pesquisa apontou que *Manhattan* é o distrito com maior preço de aluguéis de imóveis e que os bairros *Tribeca, Upper East Side, West Village e Soho* possuem valores elevados de aluguel. Dessa forma, se o registro possui uma dessas características acima, ou seja, faz parte dessas localidades, era mantido, caso contrário seria removido pois se constitui em um *outlier*. Destarte, também foram removidos os valores de preço que estavam como 0 (zero).

Posteriormente foram plotados gráficos de Histogramas para analisar a distribuição das variáveis.

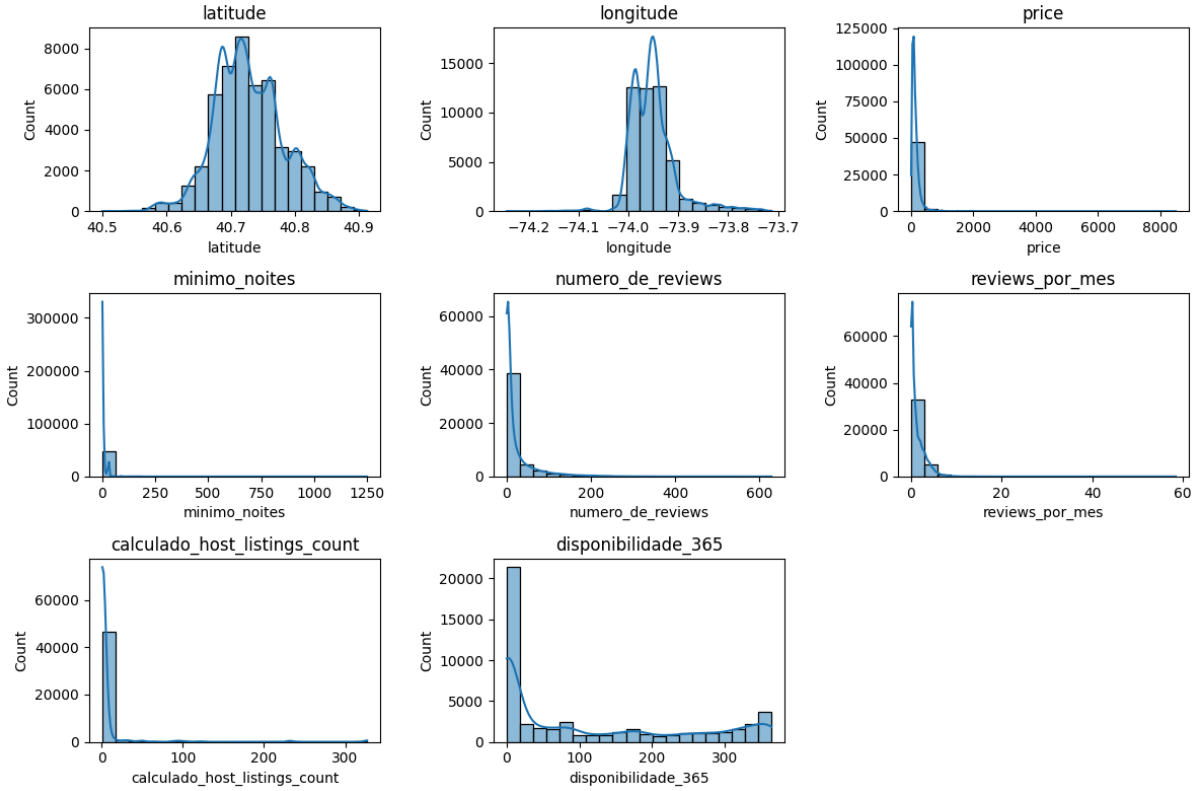


Figura 2: Histogramas das variáveis numéricas

Pode-se verificar que as únicas variáveis que se aproximam de uma distribuição normal são de latitude e longitude. As outras em sua maioria possuem uma distribuição assimétrica positiva. A transformação dessas variáveis (normalização) depende do uso do algoritmo de regressão, como foi utilizado algoritmos baseados em árvores, não foi necessário fazer a normalização.

Em seguida, a primeira forma de verificar se havia uma associação da variável objetivo (preço) com as demais variáveis numéricas foi a plotagem de gráficos de dispersão.

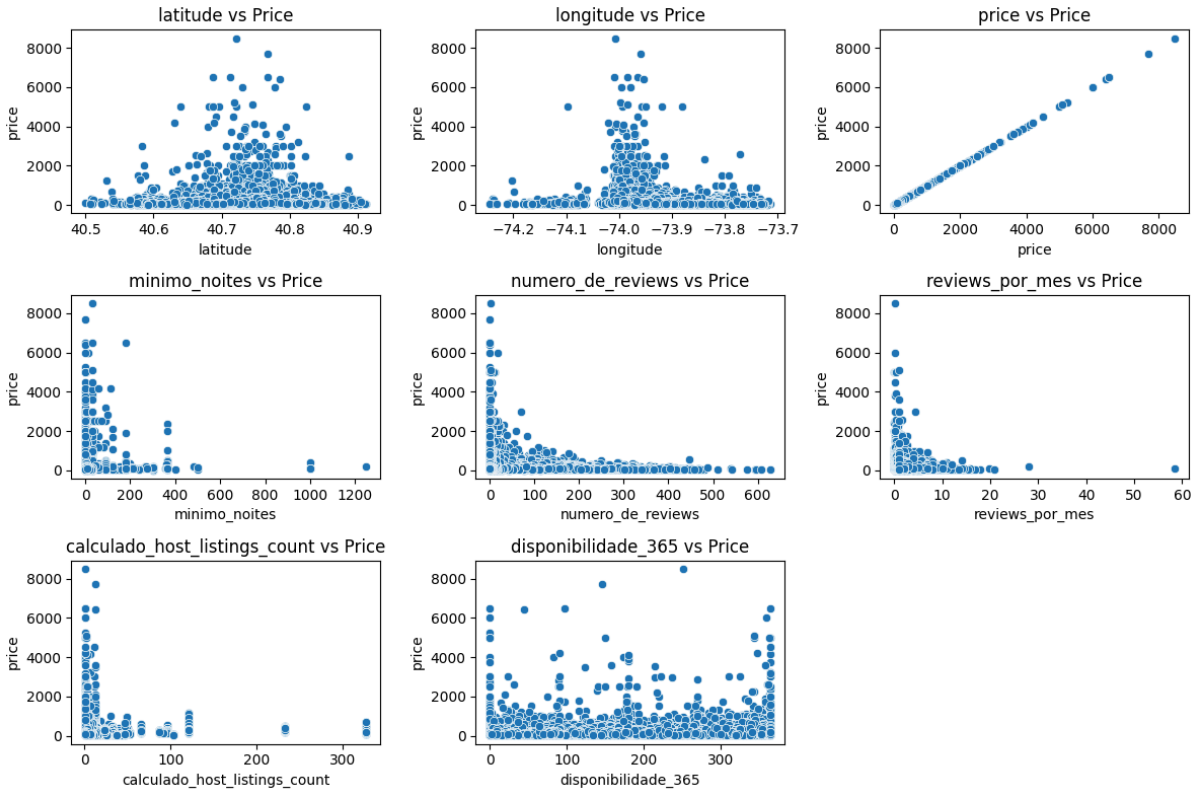


Figura 3: Gráficos de dispersão das variáveis comparadas com preço

É possível observar que as variáveis acima não possuem grande relação com a variável objetivo preço. Isso foi confirmado com os gráficos de correlação de *Pearson* e *Spearman* abaixo. Esse comportamento inicialmente gerou uma certa preocupação, uma vez que nenhuma das variáveis demonstrou uma boa associação e isso pode impactar significativamente o desempenho do modelo.

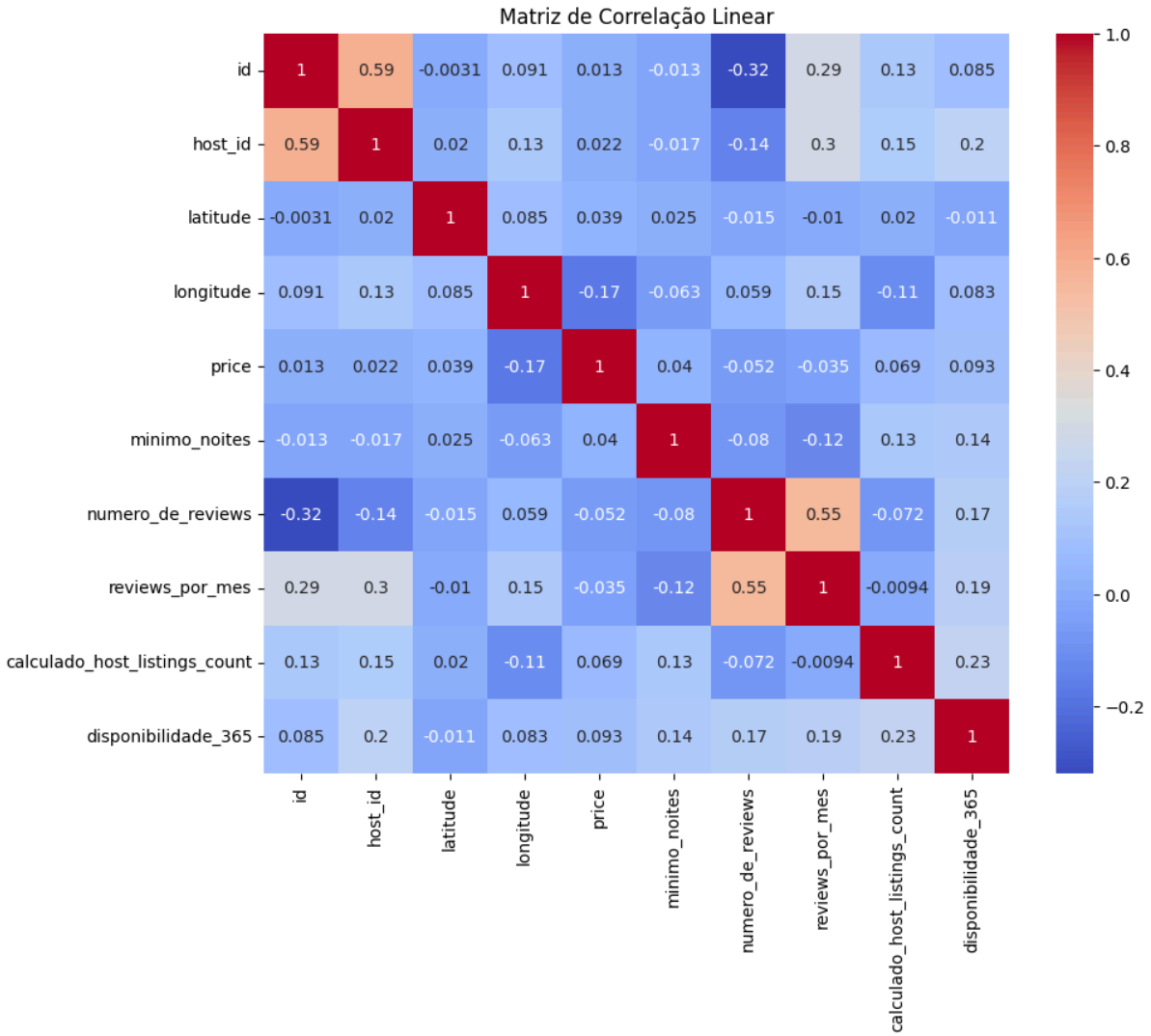


Figura 4: Gráfico de Correlação de Pearson

O gráfico de correlação de *Pearson* teve resultados inferiores ao de *Spearman* a seguir, essa correlação é linear, ou seja, indica que um algoritmo de regressão linear não deve ser eficaz para prever os preços.

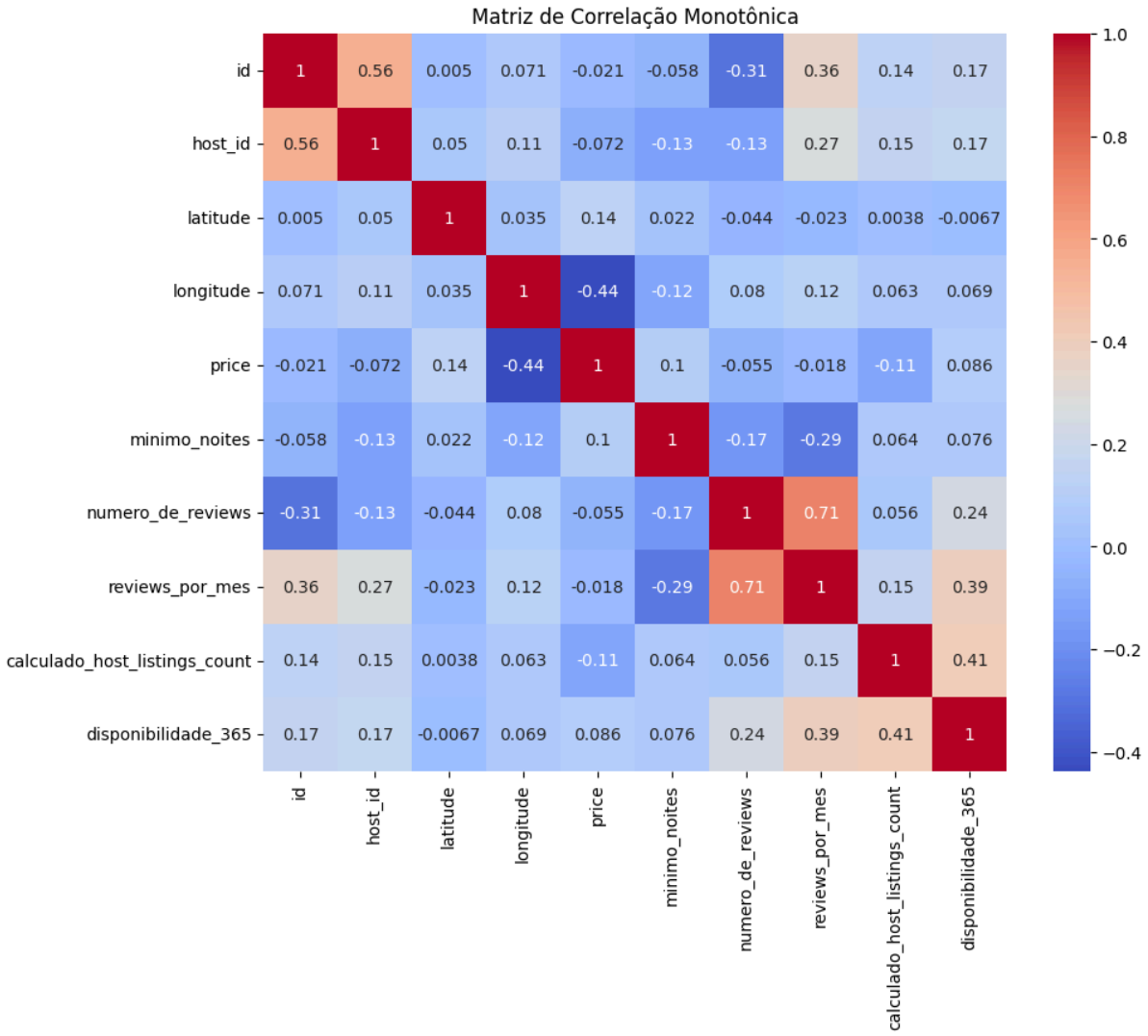


Figura 5: Gráfico de Correlação de Spearman

A correlação de Spearman teve resultados um pouco melhores do que o de Spearman, percebe-se por exemplo, que a variável longitude demonstrou uma associação negativa significante, e que houve um relativo aumento de associação em latitude e minimo_noites. Porém, as variáveis numéricas seguem não possuindo uma relação significante com o preço.

Pensando nisso, uma das perguntas do material de instrução era as variáveis de mínimo de noites e dias de disponibilidade tinham algum impacto na variável preço. A resposta está presente nos gráficos acima, o impacto é quase insignificante, mas sim, percebe-se que quanto mais os dias/noites aumentam nessa variável o preço tende a crescer ligeiramente.

5. Analisando as Variáveis Categóricas

Através da análise das variáveis categóricas é possível obter alguns insights e regras de negócio da base de dados. Com a contagem e agrupamento, bem como a média desses atributos em relação ao preço, consegue-se encontrar alguns padrões interessantes que serão mostrados a seguir.

Uma outra pergunta que deveria ser respondida no material de instruções, era se havia um padrão nos nomes de imóveis de alto valor, e sim, foi verificado que existe a repetição de algumas palavras, são elas: *Luxury*, *SuperBowl*, *Tribeca*, *Beautiful*, *Amazing*, entre outras que indicam uma localidade. Para obter esse padrão, houve a utilização do seguinte código:

```
dataframe.groupby('nome')['price'].mean().sort_values(ascending=False).head(15)
```

nome	
Beautiful/Spacious 1 bed luxury flat-TriBeCa/Soho	8500.0
East 72nd Townhouse by (Hidden by Airbnb)	7703.0
SUPER BOWL Brooklyn Duplex Apt!!	6500.0
Luxury TriBeCa Apartment at an amazing price	6500.0
Apartment New York \nHell's Kitchens	6500.0
Park Avenue Mansion by (Hidden by Airbnb)	6419.0
UWS 1BR w/backyard + block from CP	6000.0
Luxury townhouse Greenwich Village	6000.0
SuperBowl Penthouse Loft 3,000 sqft	5250.0
Midtown Manhattan great location (Gramacy park)	5100.0
4-Floor Unique Event Space 50P Cap. - #10299B	5000.0
Beautiful 1 Bedroom in Nolita/Soho	5000.0
NearWilliamsburg bridge 11211 BK	5000.0
Broadway 1	5000.0
Fulton 2	5000.0

Figura 6: Padrão de nomes nos imóveis de alto valor

Outro ponto interessante a ser analisado, foi a ocorrência de anúncio em um dos cinco distritos de Nova Iorque, sendo eles: *Manhattan, Brooklyn, Bronx, Queens e Staten Island*. No gráfico abaixo é possível notar que o distrito de *Manhattan* possui o maior número de anúncios.

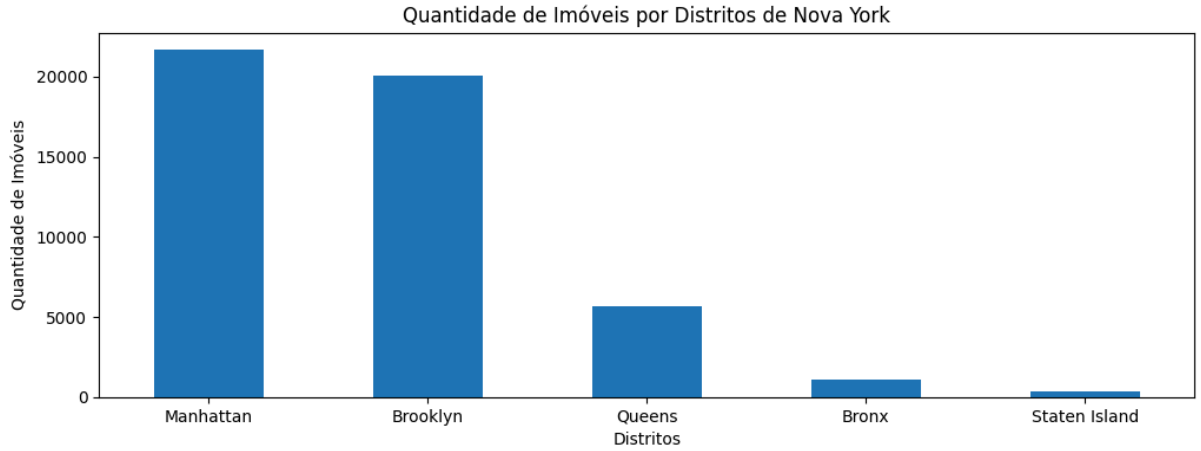


Figura 7: Quantidade de imóveis por distritos de NY

Ademais, para identificar quais os distritos eram mais caros e quais eram mais baratos, foi realizada a média de preços por distrito. O resultado mostra que *Manhattan* possui a maior média de preço entre os cinco.

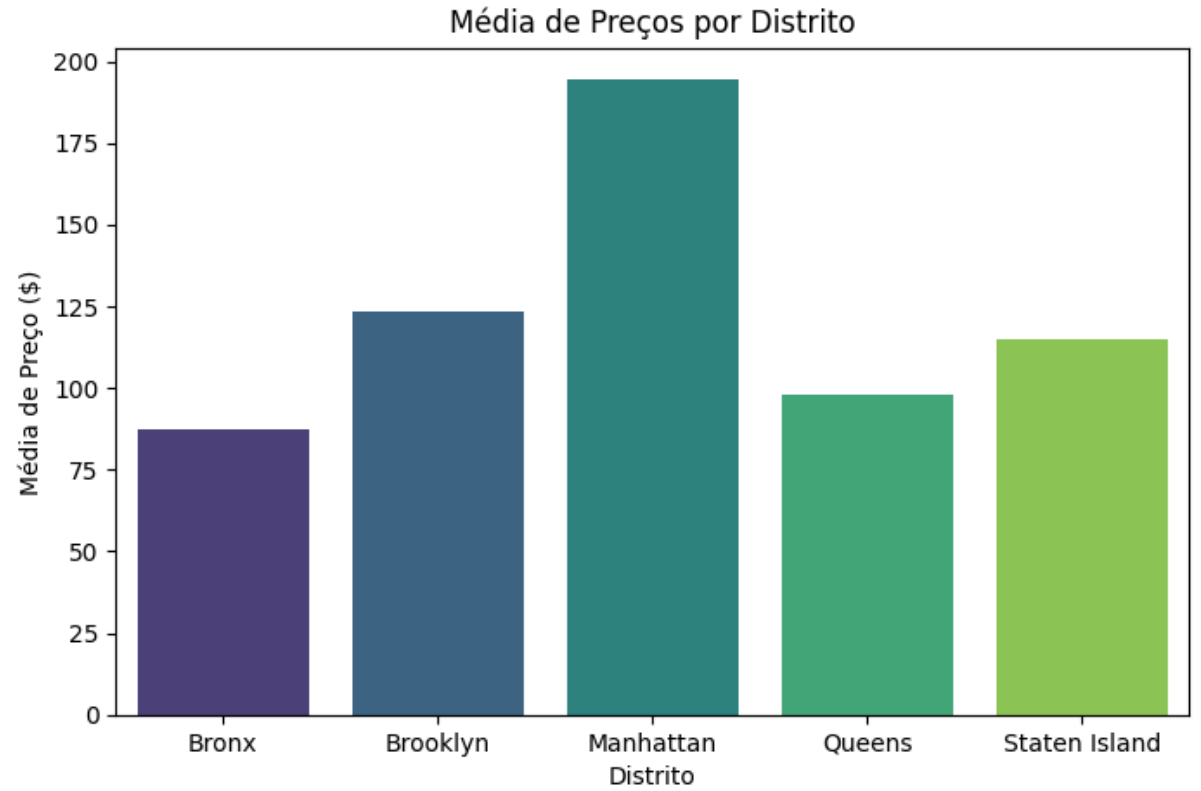


Figura 8: Média de preço por distrito de NY

A quantidade de bairros no dataset passava de 220, dessa forma, para ter uma visão parcial de quais bairros possuíam anúncios, foi feito o gráfico de maior ocorrência por bairro considerando os cinco distritos. Pode-se observar que *Williamsburg*, que fica no distrito do *Brooklyn*, é o bairro com maior quantidade de anúncios/imóveis.

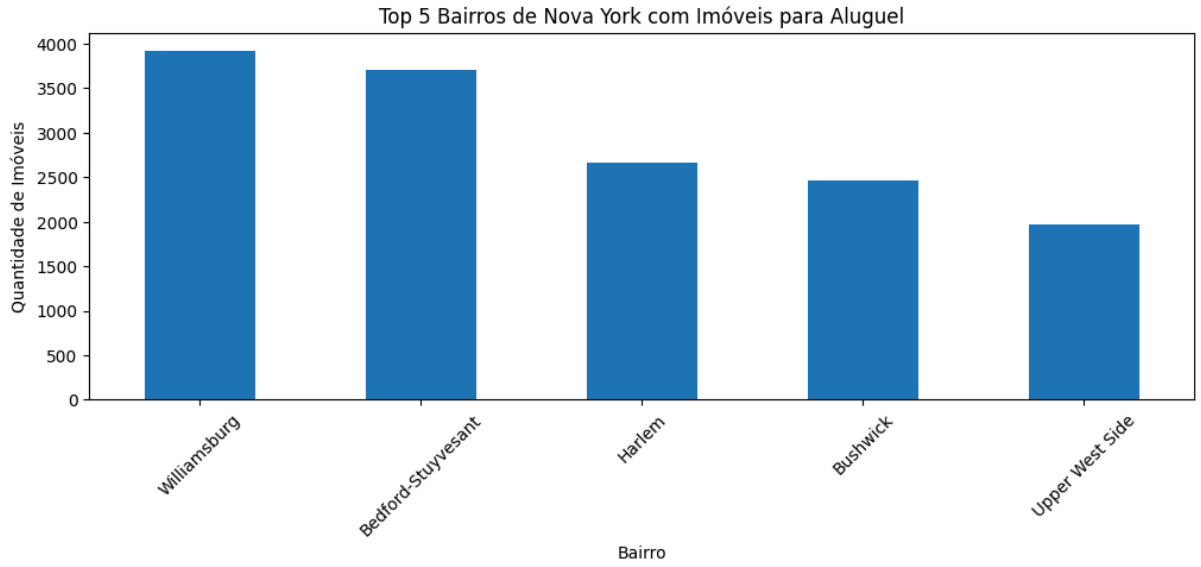


Figura 9: Bairros com maior quantidade de anúncio

Porém, por mais surpreendente, ou não, que seja, nenhum dos bairros acima estão entre os cinco com maior preço de imóveis. Os cinco bairros com maior média de preço são:

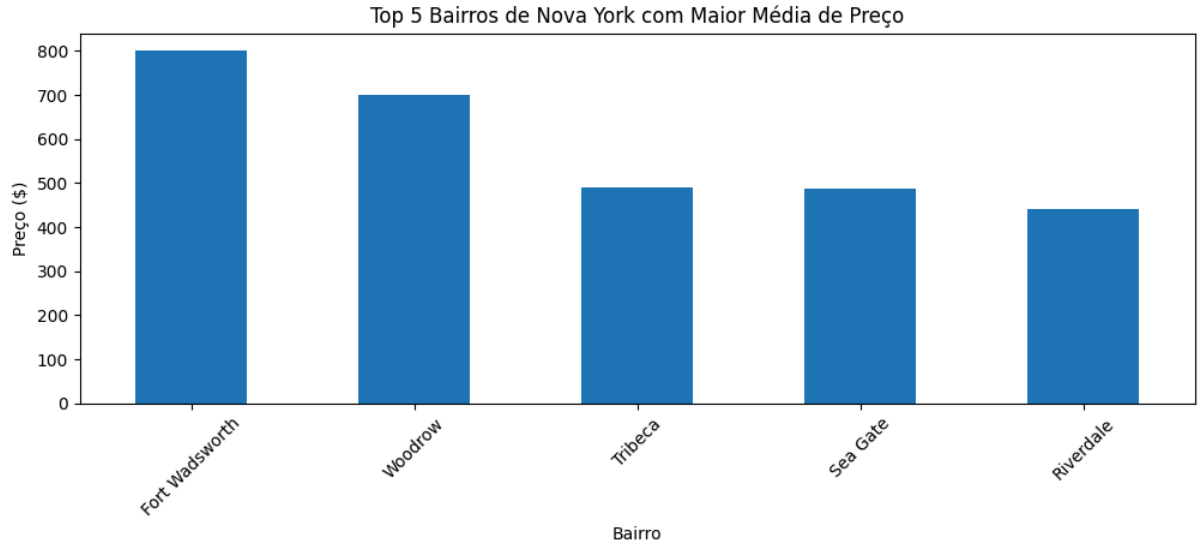


Figura 10: Bairros com maior média de preço

O bairro de *Fort Wadsworth* e *Woodrow* ficam no distrito de *Staten Island*, por sua vez o *Tribeca* fica em *Manhattan*, o *Sea Gate* fica no *Brooklyn* e *Riverdale* está localizado no *Bronx*.

Uma das variáveis mais interessantes e que tem forte impacto com o preço é a de room_type, ou seja, tipo de espaço em cada anúncio. Essa variável é dividida em 3 categorias, sendo: quarto privado, quarto compartilhado e apartamento/casa completa. Dessas categorias, a que mais aparecia nos anúncios era de apartamento/casa completo, como pode-se observar no gráfico abaixo:

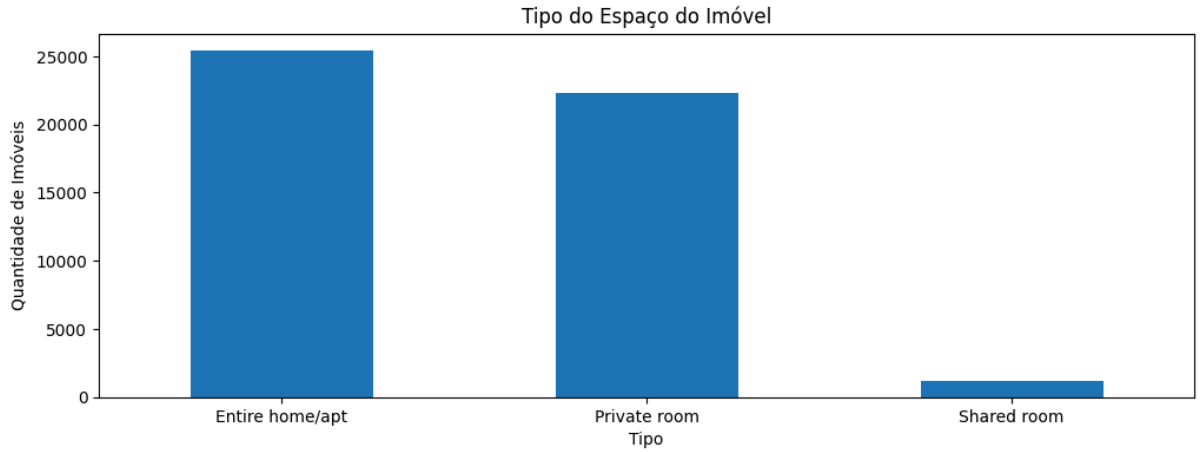
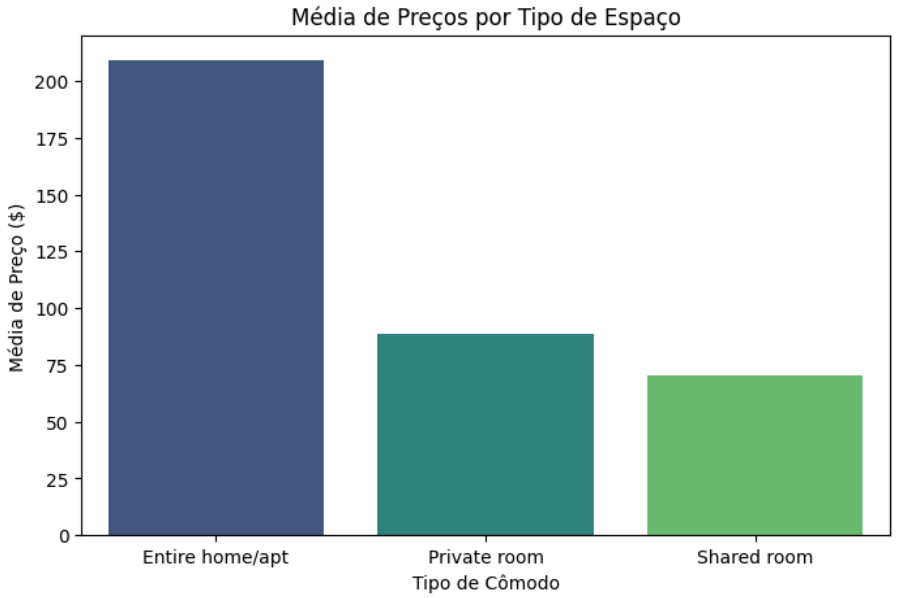


Figura 11: Quantidade de imóveis por tipo de cômodo

Outro quesito a ser analisado é o preço médio que cada uma dessas três categorias possuíam no dataset, através disso foi obtido que apartamento/casa completa também possui a maior média de preço, o que faz todo sentido, uma vez que um cômodo completo vale mais que um único cômodo.



Ademais, foi realizado um teste de ANOVA para verificar a associação das variáveis categóricas com a variável objetivo preço. Os valores do P-Valor de ANOVA para as variáveis categóricas foram:

```
nome - P-Valor da ANOVA: 7.311695396274974e-216
host_name - P-Valor da ANOVA: 4.38292497384437e-159
bairro_group - P-Valor da ANOVA: 0.0
bairro - P-Valor da ANOVA: 0.0
room_type - P-Valor da ANOVA: 0.0
ultima_review - P-Valor da ANOVA: 2.0173951198821963e-262
```

Baseado nos valores acima, infere-se que apenas as variáveis *bairro*, *bairro_group* e *room_type* possuem uma relação significativa com a variável objetivo preço. Isso significa que o preço varia bastante dependendo da localização e do tipo de acomodação.

6. Valores Nulos e Duplicados

Para fazer a verificação da existência dos valores nulos e duplicados, foi utilizado dois métodos, sendo *duplicated()* para verificar dados duplicados e *isnull().sum()* para verificar a quantidade de valores faltantes.

Através da chamada do método acima, não foram encontrados registros duplicados no dataset, contudo foram encontrados valores nulos em algumas variáveis, sendo elas: *nome, host_name, ultima_review, reviews_por_mes*. Dessas variáveis estarei utilizando apenas a *ultima_review* para fazer uma engenharia de atributos criando uma nova variável através dela. Para isso, preenchi os valores faltantes da variável com 0 (zero).

7. Conclusão

Infere-se, portanto, que a etapa de análise exploratória dos dados é uma das técnicas mais importantes ao fazer previsões com machine learning. Isso acontece, visto que é nessa etapa que conhecemos o dataset e conseguimos fazer conexões entre as variáveis para obter insights. A título de exemplo, os gráficos de bairros, distritos e tipos de espaços com maior média de preço foi obtido através do agrupamento e comparação com a variável preço, desse modo, pode-se perceber que os distritos com maiores preço são por exemplo, Manhattan e Brooklyn e que os tipos de espaço com maiores preços são os imóveis completos. Dessa forma, a EDA mostra-se como a melhor forma de entender o seu conjunto de dados.