analise_eda

January 31, 2025

1 1. Análise Exploratória dos Dados (EDA)

1.1 1.1: Carregando os dados

[3]:
```
# Importando as bibliotecas necessárias para a análise
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[4]:
```
# Carregando a base de dados
df_precificacao = pd.read_csv('teste_indicium_precificacao.csv')
```

[5]:
```
# Visualizando as primeiras linhas
df_precificacao.head()
```

[5]:

	id	nome	host_id	\
0	2595	Skylit Midtown Castle	2845	
1	3647	THE VILLAGE OF HARLEM…NEW YORK !	4632	
2	3831	Cozy Entire Floor of Brownstone	4869	
3	5022	Entire Apt: Spacious Studio/Loft by central park	7192	
4	5099	Large Cozy 1 BR Apartment In Midtown East	7322	

	host_name	bairro_group	bairro	latitude	longitude	\
0	Jennifer	Manhattan	Midtown	40.75362	-73.98377	
1	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	
2	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	
3	Laura	Manhattan	East Harlem	40.79851	-73.94399	
4	Chris	Manhattan	Murray Hill	40.74767	-73.97500	

	room_type	price	minimo_noites	numero_de_reviews	ultima_review	\
0	Entire home/apt	225	1	45	2019-05-21	
1	Private room	150	3	0	NaN	
2	Entire home/apt	89	1	270	2019-07-05	
3	Entire home/apt	80	10	9	2018-11-19	
4	Entire home/apt	200	3	74	2019-06-22	

	reviews_por_mes	calculado_host_listings_count	disponibilidade_365

```
0              0.38                                2                  355
1               NaN                                1                  365
2              4.64                                1                  194
3              0.10                                1                    0
4              0.59                                1                  129
```

[6]:
```
# Visualizando as últimas linhas
df_precificacao.tail()
```

[6]:
```
             id                                                 nome   host_id  \
48889  36484665    Charming one bedroom - newly renovated rowhouse   8232441
48890  36485057      Affordable room in Bushwick/East Williamsburg   6570630
48891  36485431            Sunny Studio at Historical Neighborhood  23492952
48892  36485609               43rd St. Time Square-cozy single bed  30985759
48893  36487245  Trendy duplex in the very heart of Hell's Kitchen  68119814

          host_name bairro_group              bairro  latitude  longitude  \
48889       Sabrina     Brooklyn  Bedford-Stuyvesant  40.67853  -73.94995
48890       Marisol     Brooklyn            Bushwick  40.70184  -73.93317
48891  Ilgar & Aysel    Manhattan              Harlem  40.81475  -73.94867
48892           Taz    Manhattan      Hell's Kitchen  40.75751  -73.99112
48893     Christophe    Manhattan      Hell's Kitchen  40.76404  -73.98933

            room_type  price  minimo_noites  numero_de_reviews ultima_review  \
48889     Private room     70              2                  0           NaN
48890     Private room     40              4                  0           NaN
48891  Entire home/apt    115             10                  0           NaN
48892      Shared room     55              1                  0           NaN
48893     Private room     90              7                  0           NaN

       reviews_por_mes  calculado_host_listings_count  disponibilidade_365
48889              NaN                              2                    9
48890              NaN                              2                   36
48891              NaN                              1                   27
48892              NaN                              6                    2
48893              NaN                              1                   23
```

[7]:
```
# Informações gerais
df_precificacao.info()
```

```
<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
```

[8]:
```
# Estatísticas descritivas
df_precificacao.describe()
```

[8]:

	id	host_id	latitude	longitude	price
count	4.889400e+04	4.889400e+04	48894.000000	48894.000000	48894.000000
mean	1.901753e+07	6.762139e+07	40.728951	-73.952169	152.720763
std	1.098288e+07	7.861118e+07	0.054529	0.046157	240.156625
min	2.595000e+03	2.438000e+03	40.499790	-74.244420	0.000000
25%	9.472371e+06	7.822737e+06	40.690100	-73.983070	69.000000
50%	1.967743e+07	3.079553e+07	40.723075	-73.955680	106.000000
75%	2.915225e+07	1.074344e+08	40.763117	-73.936273	175.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000

	minimo_noites	numero_de_reviews	reviews_por_mes
count	48894.000000	48894.000000	38842.000000
mean	7.030085	23.274758	1.373251
std	20.510741	44.550991	1.680453
min	1.000000	0.000000	0.010000
25%	1.000000	1.000000	0.190000
50%	3.000000	5.000000	0.720000
75%	5.000000	24.000000	2.020000
max	1250.000000	629.000000	58.500000

	calculado_host_listings_count	disponibilidade_365
count	48894.000000	48894.000000
mean	7.144005	112.776169
std	32.952855	131.618692
min	1.000000	0.000000
25%	1.000000	0.000000
50%	1.000000	45.000000
75%	2.000000	227.000000

```
max                            327.000000            365.000000
```

1.2 1.2: Limpeza dos dados

[10]:
```
# Verificando quantidade de valores faltantes em cada coluna
df_precificacao.isnull().sum()
```

[10]:
```
id                                  0
nome                               16
host_id                             0
host_name                          21
bairro_group                        0
bairro                              0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimo_noites                       0
numero_de_reviews                   0
ultima_review                   10052
reviews_por_mes                 10052
calculado_host_listings_count       0
disponibilidade_365                 0
dtype: int64
```

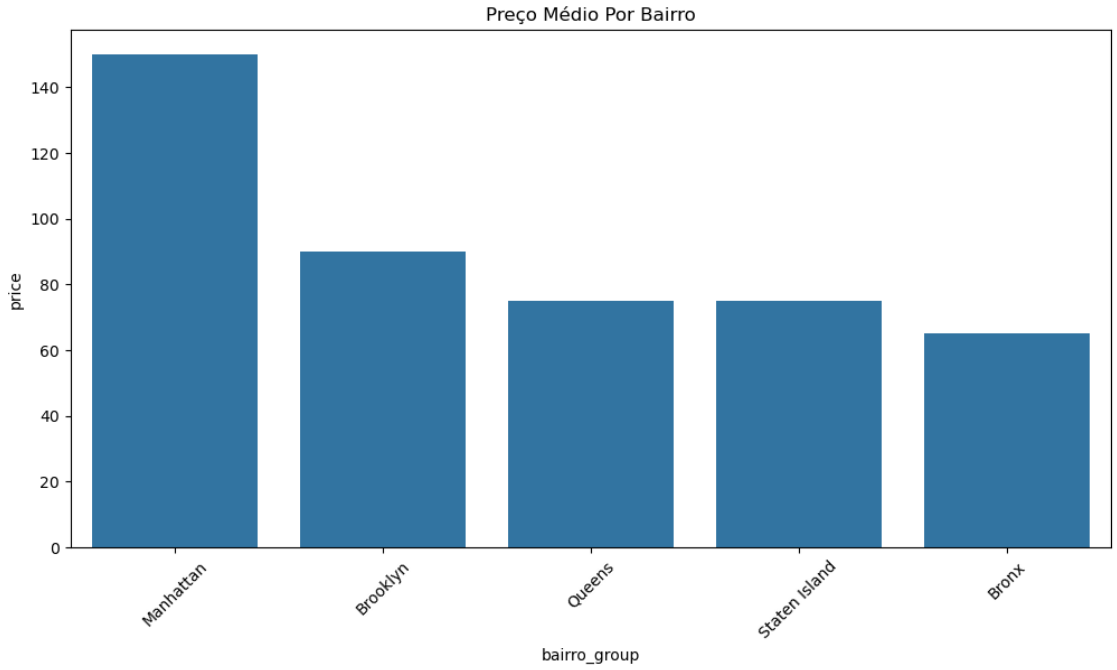
[11]:
```
# Removendo colunar com mais de 30% de dados faltando
limite = len(df_precificacao) * 0.7
df_precificacao = df_precificacao.dropna(thresh=limite, axis=1)
```

[12]:
```
# Removendo duplicatas baseadas no id
df_precificacao = df_precificacao.drop_duplicates(subset=['id'])
```

1.3 1.3: Estabelecendo hipóteses

1.3.1 Hipótese 1: Os bairros localizados em Manhattan têm preços mais elevados.

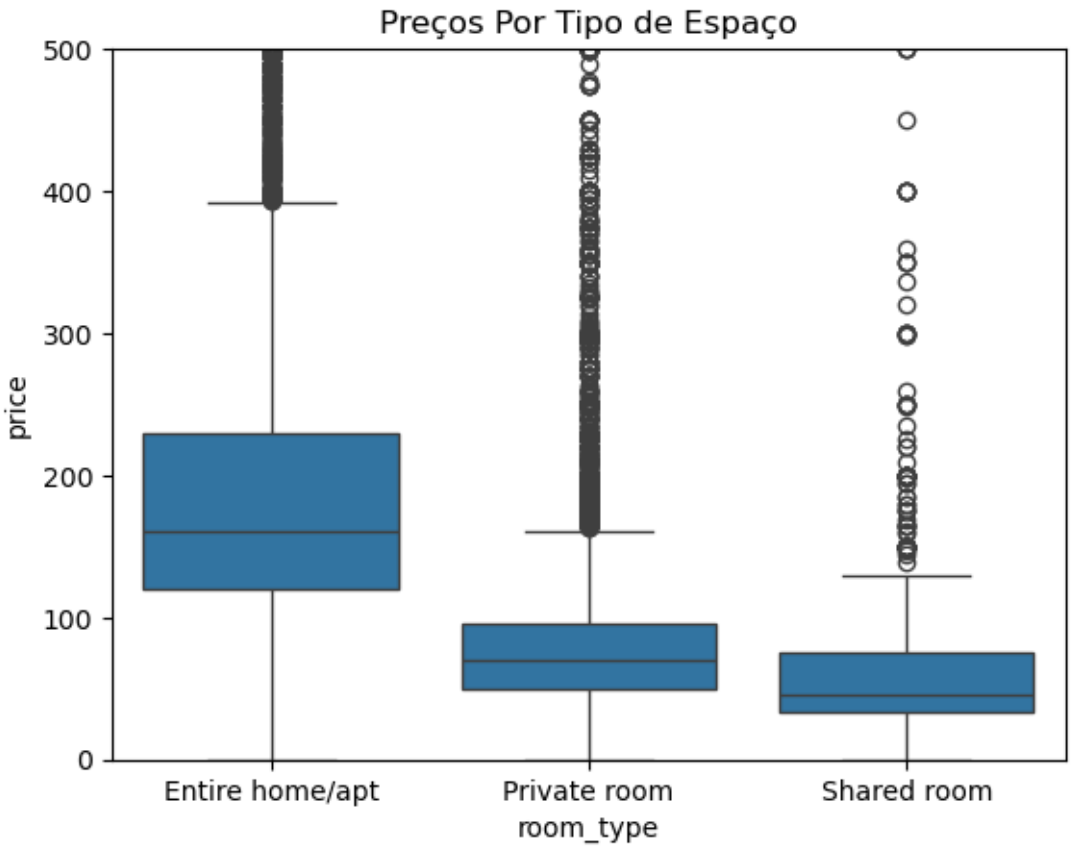
[15]:
```
# Calculando o preço médio por bairro_group (Top 10) e armazenando em um novo␣
 ↪dataframe
df_pmbg = df_precificacao.groupby('bairro_group', as_index=False)['price'].
 ↪median().sort_values('price', ascending=False).head(10)

# Criando um gráfico em barras para visualização dos resultados acima
plt.figure(figsize=(12, 6))
sns.barplot(df_pmbg, x='bairro_group', y='price')
plt.xticks(rotation=45)
plt.title('Preço Médio Por Bairro')
plt.show()
```



1. O gráfico mostra que Manhattan possui o preço médio mais elevado entre os analisados. Isso confirma a hipótese que bairros localizados nessa região têm preços mais caros.
2. Há uma disparidade considerável de valores entre Manhattan e outras regiões, indicando uma valorização das propriedades nesse local.
3. Alguns fatores contribuintes para a confirmação da hipótese são: oferta de terreno e imoveis mais limitada (por ser uma ilha), centro econômico global, turismo e cultura, etc.

1.3.2 Hipótese 2: Espaços do tipo casa inteira ou apartamento (Entire home/apt) são mais caros.

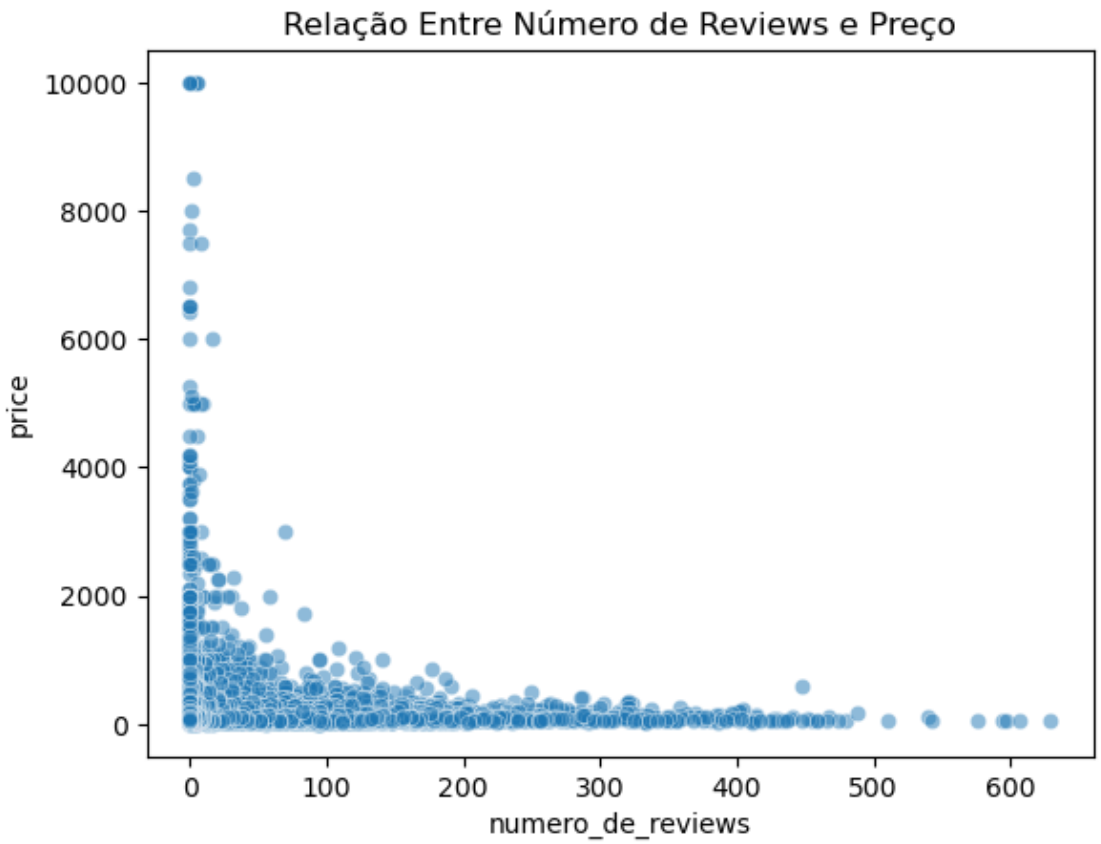
[18]:
```
# Criando um gráfico em boxplot por tipo de espaço
sns.boxplot(df_precificacao, x='room_type', y='price')
plt.ylim(0, 500) # Remove outliers
plt.title('Preços Por Tipo de Espaço')
plt.show()
```



1. Os dados confirmam a hipótese de que espaços do tipo casa inteira ou apartamento (Entire home/apt) são mais caros.
2. Espaços do tipo Entire home/apt têm mediana de preço mais alta em comparação aos outros espaços.
3. Variação maior de preços em Entire home/apt. Provavelmente por diferenças de tamanho, localização e comodidades oferecidas.
4. Alguns outliers nas 3 cateorias. Maior quantidade em Entire home/apt, sugerindo que alguns espaços deste tipo podem ser excepcionalmente caros.

1.3.3 Hipótese 3: Listagens com maior número de reviews têm preços mais elevados.

[21]:
```
# Criando um gráfico de dispersão com correlação entre o número de reviews e␣
↪preço
sns.scatterplot(df_precificacao, x='numero_de_reviews', y='price', alpha=0.5)
plt.title('Relação Entre Número de Reviews e Preço')
plt.show()
```



1. As informações não confirmam a hipótese em questão.
2. O gráfico demonstra que o número de reviews não é um forte indicador do preço.
3. Listagens com preços menores variam mais a quantidade de reviews, enquanto preços mais elevados tendem a ter menos reviews.

2 2. Respondendo algumas perguntas

2.1 2.1: Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

[25]:
```
# Calculando o preço médio e demanda (baseada em reviews) por bairro
df_analise_bairro = df_precificacao.groupby('bairro').agg(preco_medio=('price',␣
↪'median'), demanda=('numero_de_reviews', 'sum')).sort_values('demanda',␣
↪ascending=False).head(10)

df_analise_bairro
```

[25]:

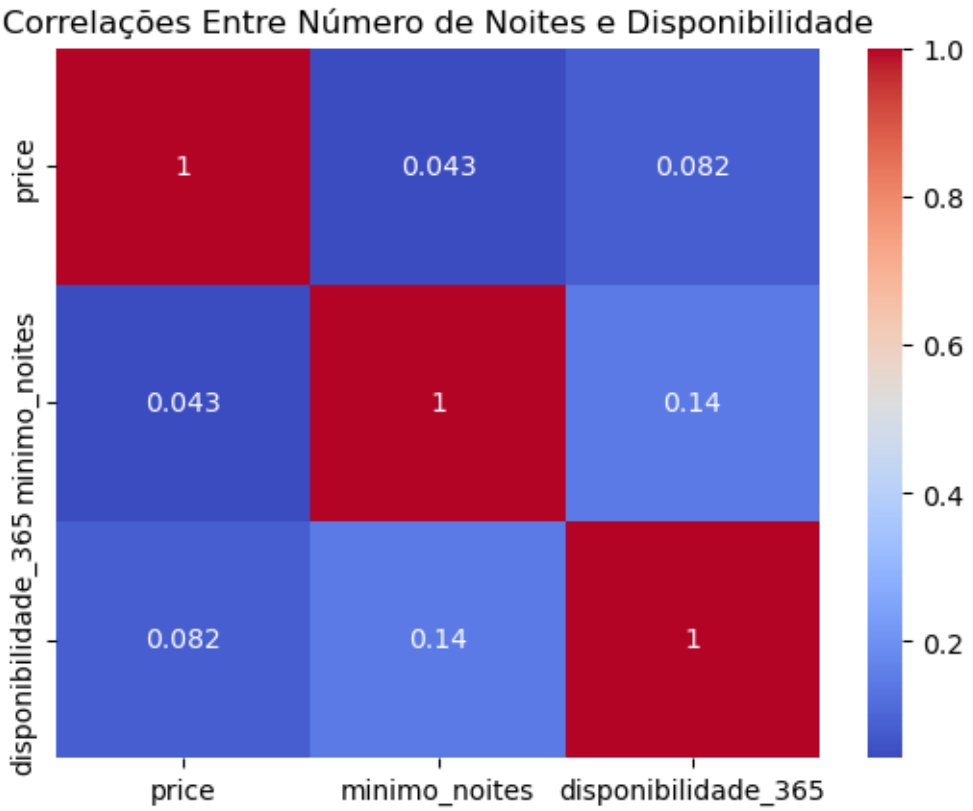
bairro	preco_medio	demanda
Bedford-Stuyvesant	80.0	110352

Williamsburg	105.0	85427
Harlem	89.0	75962
Bushwick	65.0	52514
Hell's Kitchen	168.0	50227
East Village	150.0	44670
East Harlem	99.0	36446
Crown Heights	85.0	36408
Upper West Side	150.0	36058
Upper East Side	149.0	31686

Resposta: Conforme os dados, **Bedford-Stuyvesant** e **Williamsburg** são os bairros mais recomendados para o investimento, ao apresentam uma demanda mais alta e preços médios mais competitivos, trazendo um equilíbrio entre estes fatores.

2.2 2.2: O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

[28]:
```
# Verificando a correlação número mínimo de noites e disponibiliade no ano
df_corr_noites_disponibilidade = df_precificacao[['price', 'minimo_noites',␣
 ↪'disponibilidade_365']].corr()
sns.heatmap(df_corr_noites_disponibilidade, annot=True, cmap='coolwarm')
plt.title('Correlações Entre Número de Noites e Disponibilidade')
plt.show()
```

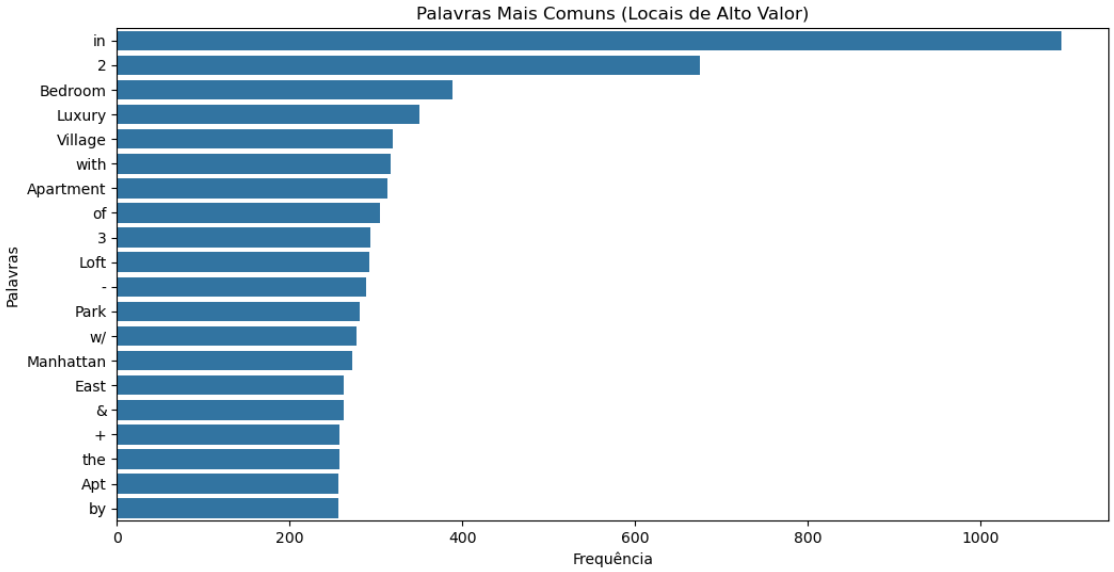


Resposta: 1. **Correlação entre número mínimo de noites e preço:** Aproximadamente de 0.043, indicando uma correlação bastante fraca e positiva, o que sugere que o mínimo de noites não gera impacto significativo no preço. 2. **Correlação entre disponibilidade e preço:** Aproximadamente de 0.082, indicando também uma correlação muito fraca e positiva e portanto que a disponibilidade ao longo do ano não gera impacto no preço.

Portanto, nenhum dos 2 fatores parece interferir no preço das listagens.

2.3 2.3: Existe algum padrão no texto do nome do local para lugares de mais alto valor?

[31]:
```
# Filtrando os 10% locais mais caros
df_top10_caros = df_precificacao[df_precificacao['price'] >=␣
 ↪df_precificacao['price'].quantile(0.9)]

# Contando as palavras no nome dos locais filtrados
contagem_palavras = df_top10_caros['nome'].str.split(expand=True).stack().
 ↪value_counts()

# Coletando as 20 palavras mais comuns
palavras = contagem_palavras.head(20)

# Demonstrando em um gráfico de barras
plt.figure(figsize=(12, 6))
sns.barplot(x=palavras.values, y=palavras.index, orient='h')
plt.xlabel('Frequência')
plt.ylabel('Palavras')
plt.title('Palavras Mais Comuns (Locais de Alto Valor)')
plt.show()
```



Resposta:

1. Termos como “Village, Park, Manhattan” e “East” são frequentes, indicando que um fator importante para os locais de alto valor seria a localização.
2. Palavras como “Bedroom, Luxury, Apartment” e “Loft” também aparecem com frequência, sugerindo que estas são destaque nas descrições das listagens de alto valor.
3. Outros termos como “with, w/, by” podem indicar espaços com características adicionais como “with pool” ou “by park”.