

Relatório de análises estatísticas

Jonatas Halliday S. Nascimento

Todos os gráficos produzidos nesta análise foram feitos em Jupyter Notebook. Para melhor especificação, sugiro conferir o notebook.

O notebook é dividido em 3 partes principais, sendo elas:

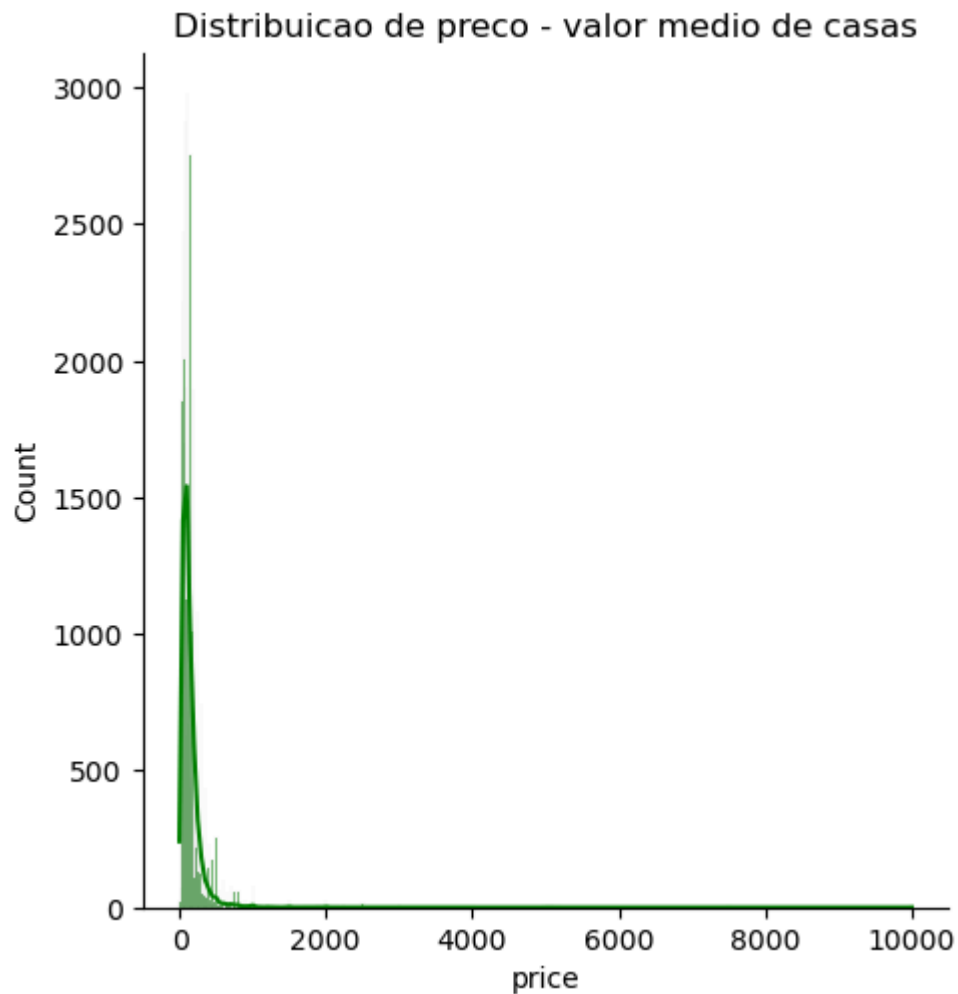
- 1) Carregamento de bibliotecas necessárias
- 2) Análise e limpeza de dados
- 3) Exercício de predição

Carregamento de bibliotecas necessárias

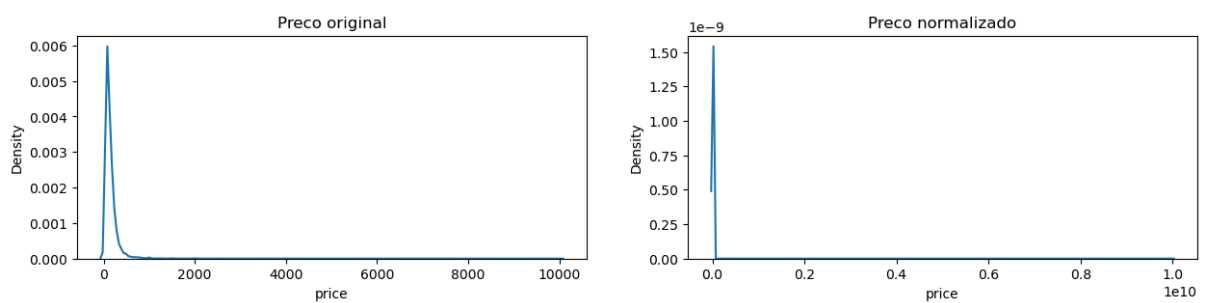
Nesta seção ocorre apenas o *import* das bibliotecas necessárias para o uso da execução completa do projeto.

Análise e limpeza de dados

Nesta seção ocorre todo tipo de limpeza julgada necessária para responder às questões propostas no desafio e para o melhor cumprimento de todas as expectativas. Inicialmente para ter uma ideia do todo, foi feito o gráfico de distribuição do preço de todo o *dataset* conforme imagem abaixo.



Após isso, como hipóteses, percebi uma necessária normalização da curva para fazer com que ela se aproxime o máximo possível da distribuição normal.

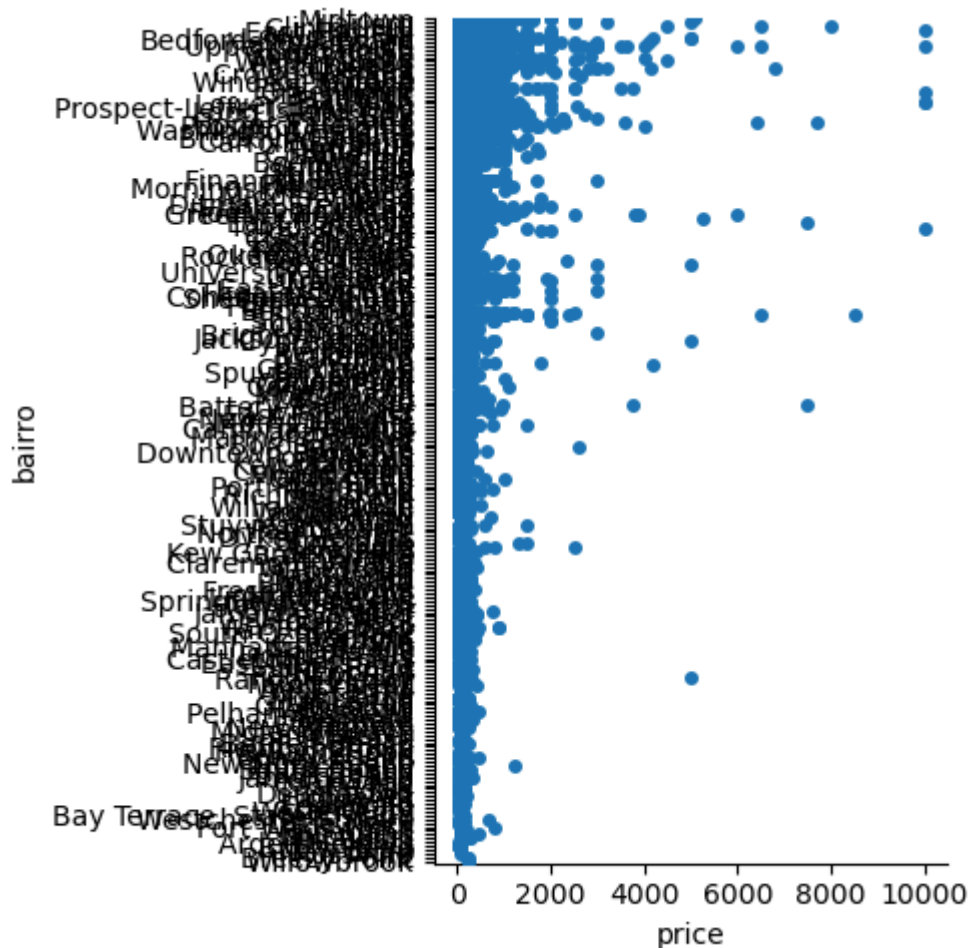


Após conferir os gráficos lado a lado, tentei outra abordagem procurando responder diretamente às perguntas propostas.

Q1: Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

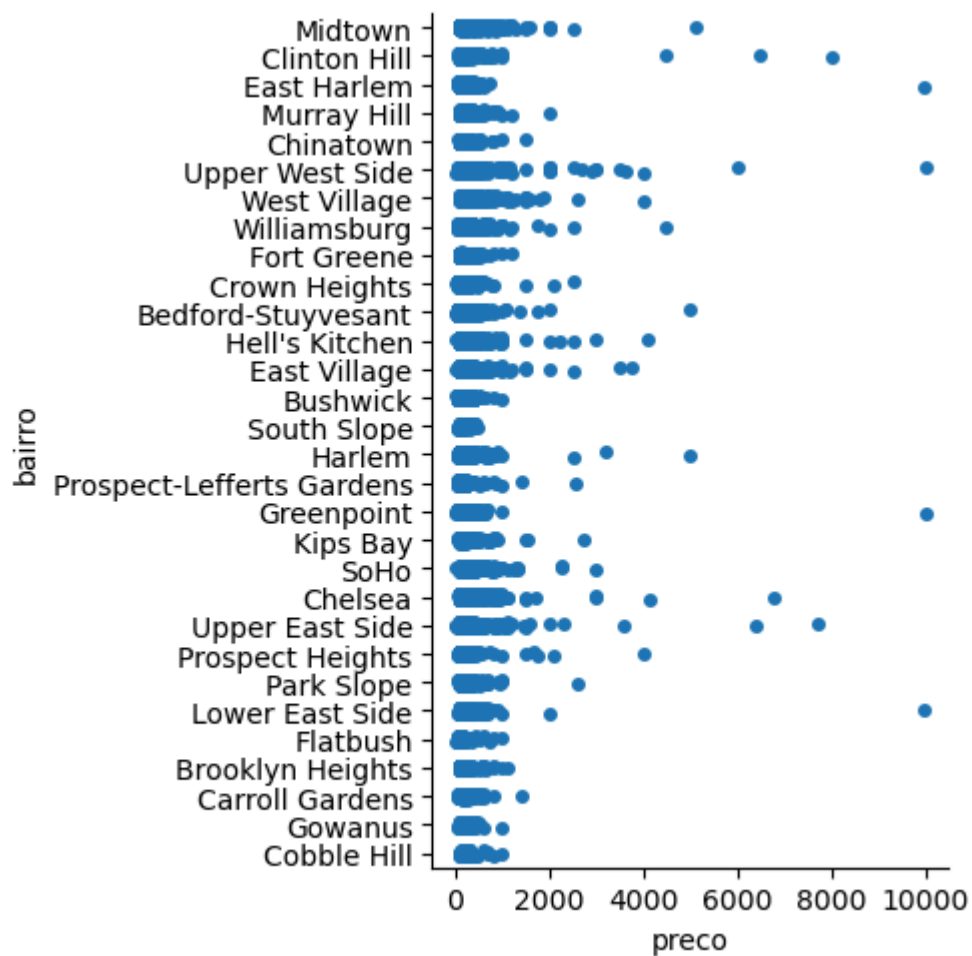
Para responder a pergunta, parti da hipótese de que o local mais indicado para compra seria o mais barato e uma vez que deseja-se uma compra, procurei pelo tipo 'Entire home/apt' uma vez que não dá para comprar um quarto, por exemplo. Basicamente fiz um filtro por imóveis(seja ele apartamento ou casa).

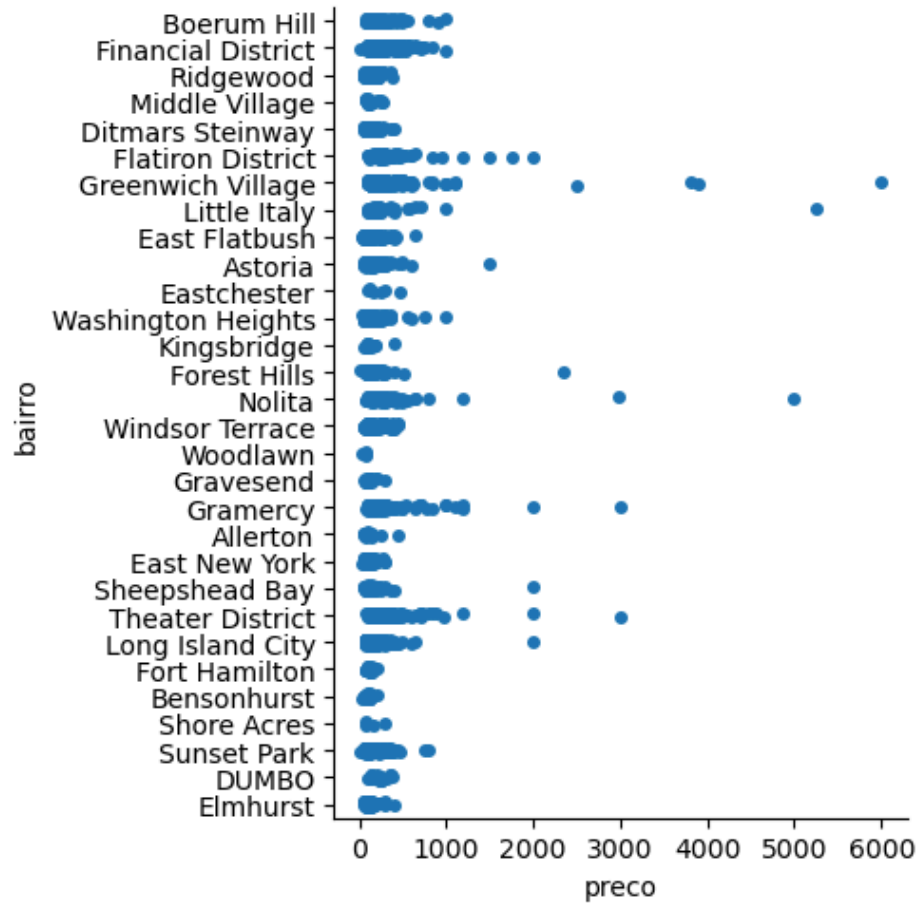
Para uma melhor visualização, tentei construir o gráfico na forma *scatter plot* para observar a distribuição e notei que há muitos dados, e os *labels* acabaram ficando sobrepostos.

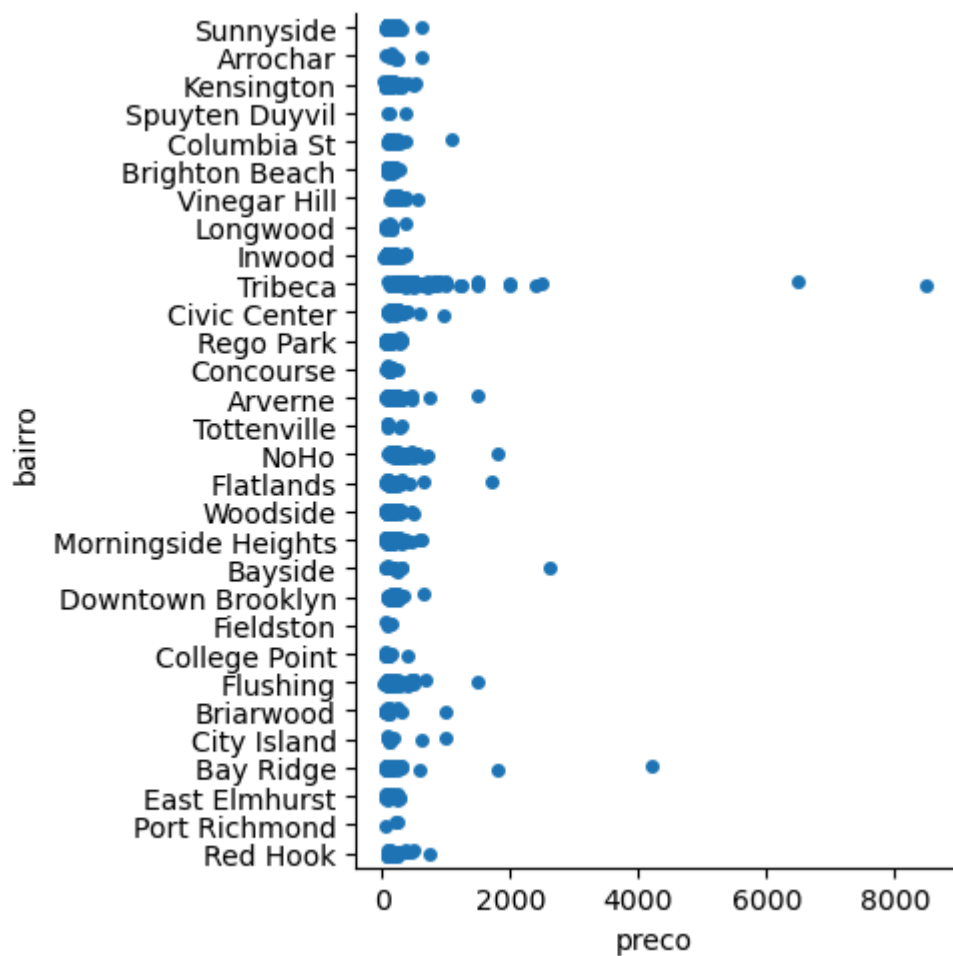


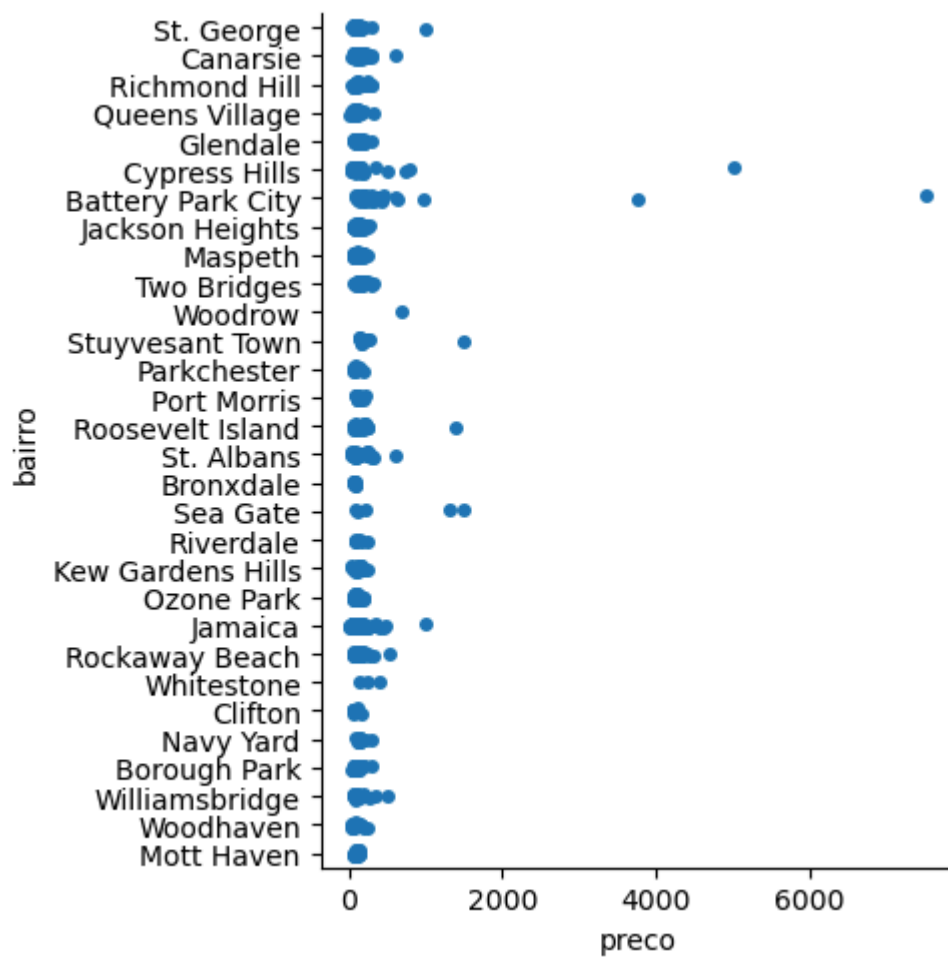
Para contornar, selecionei apenas os tipos 'Entire home/apt' da coluna *room_type* e subdivi em listas de 30 bairros.

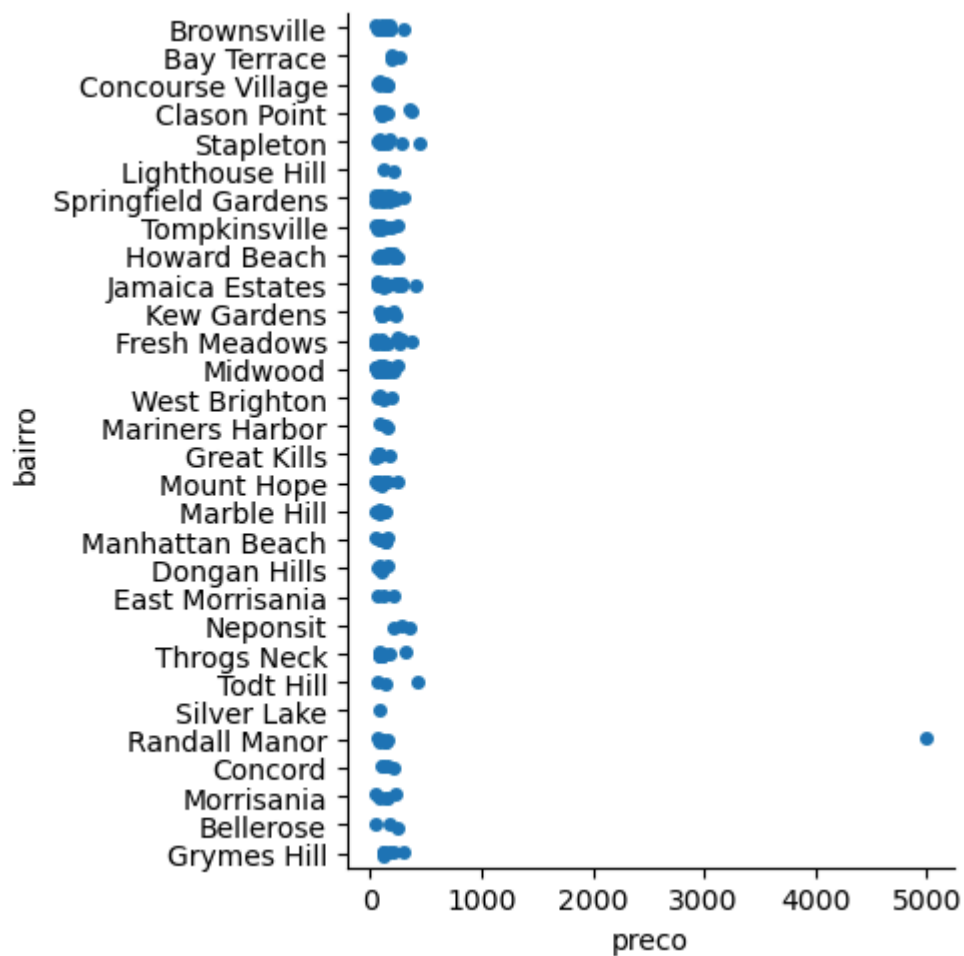
Após isso foi feito os gráficos com essas subdivisões, para finalmente ter uma ideia. Foram eles:

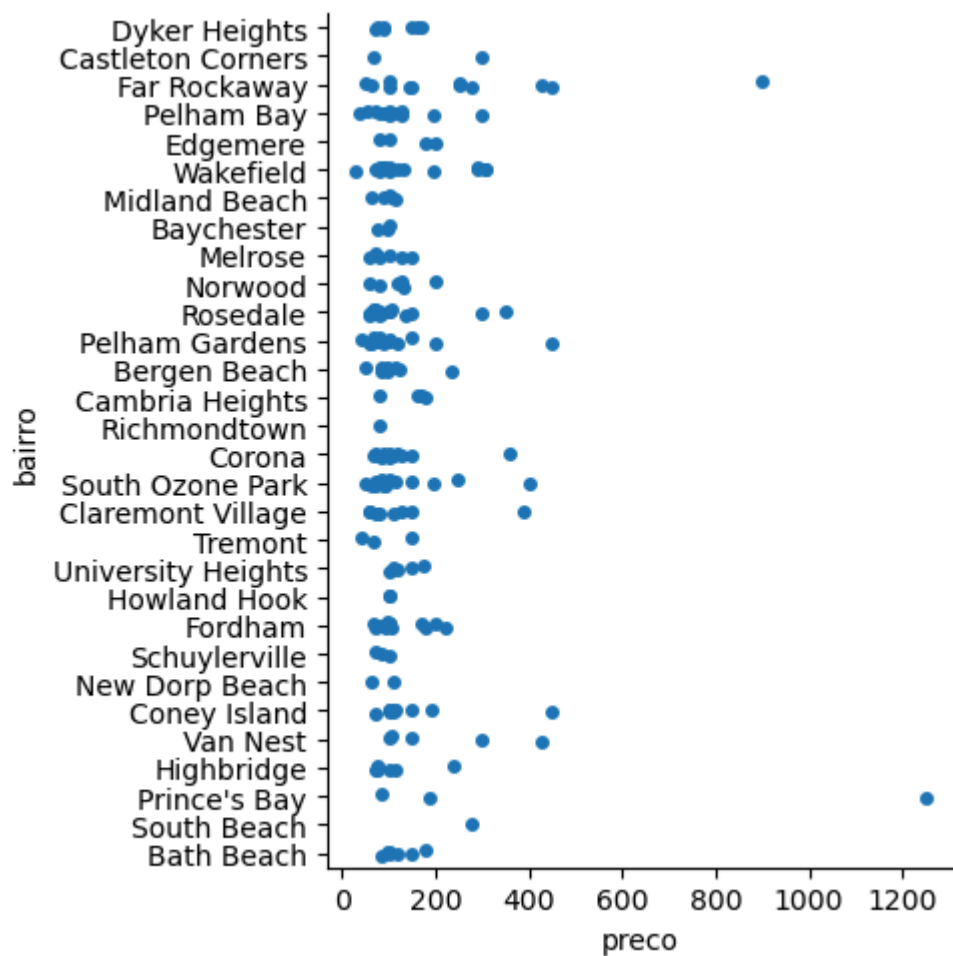


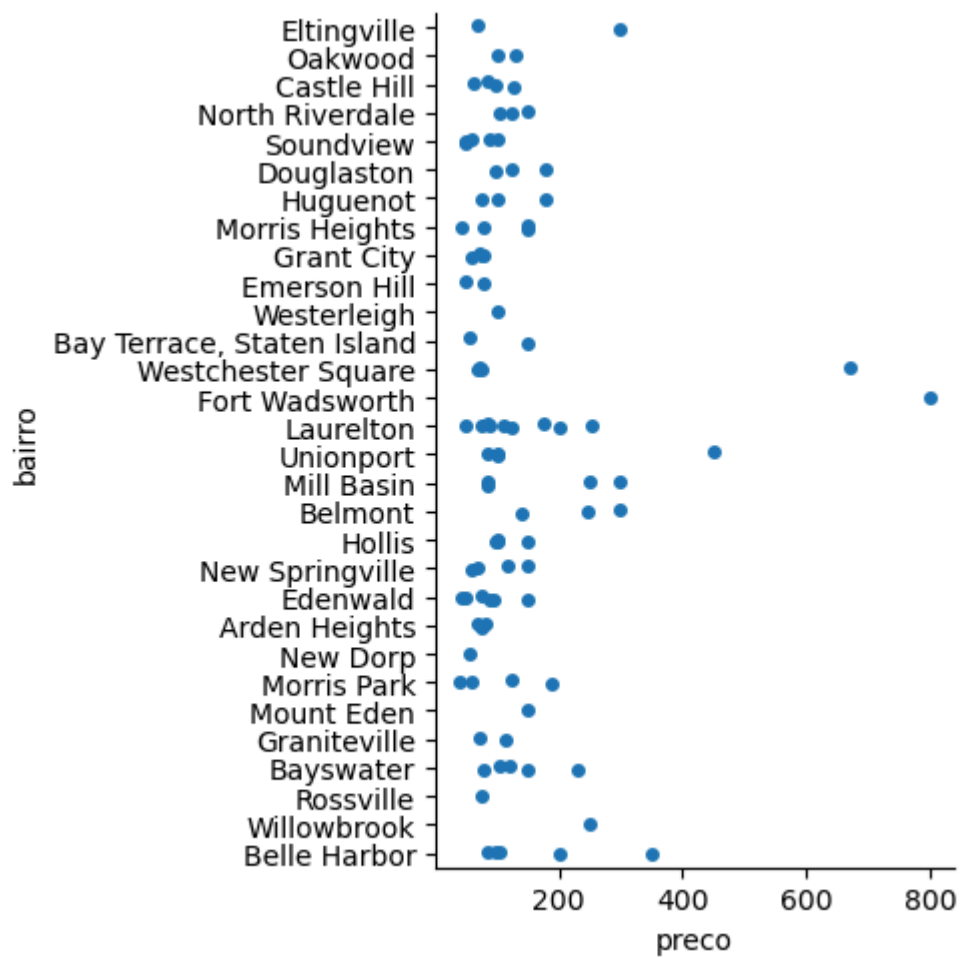


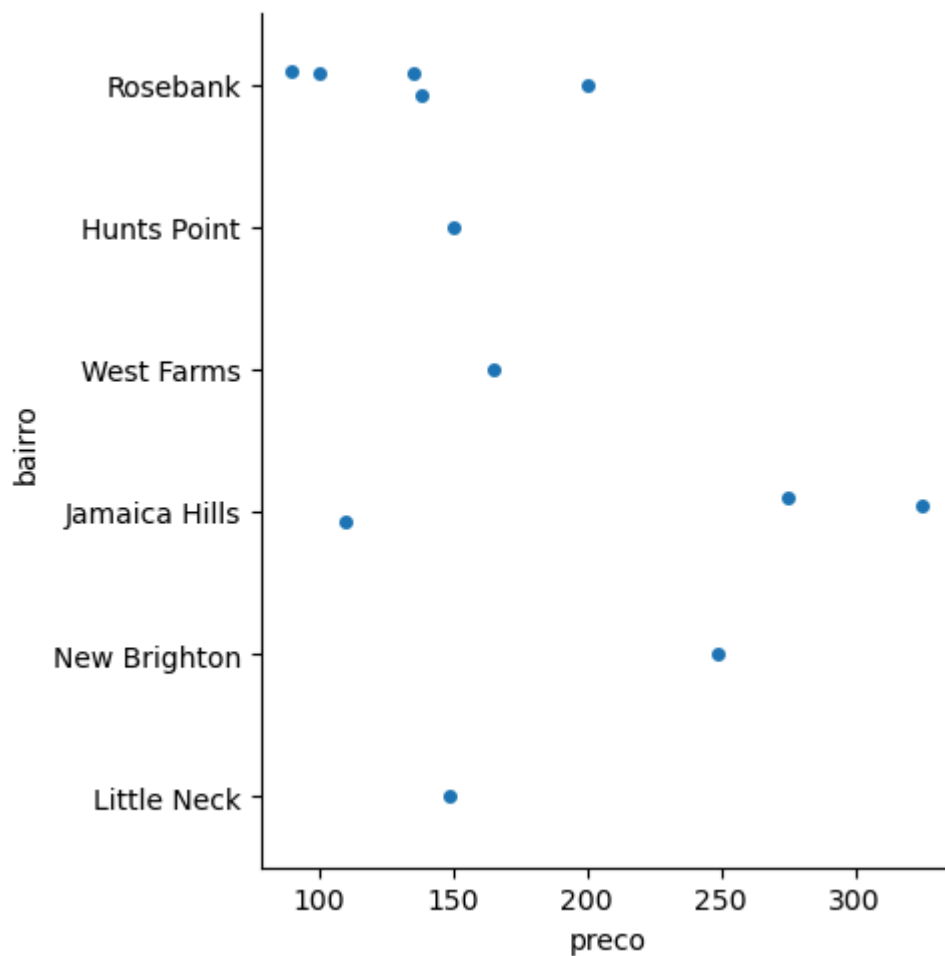












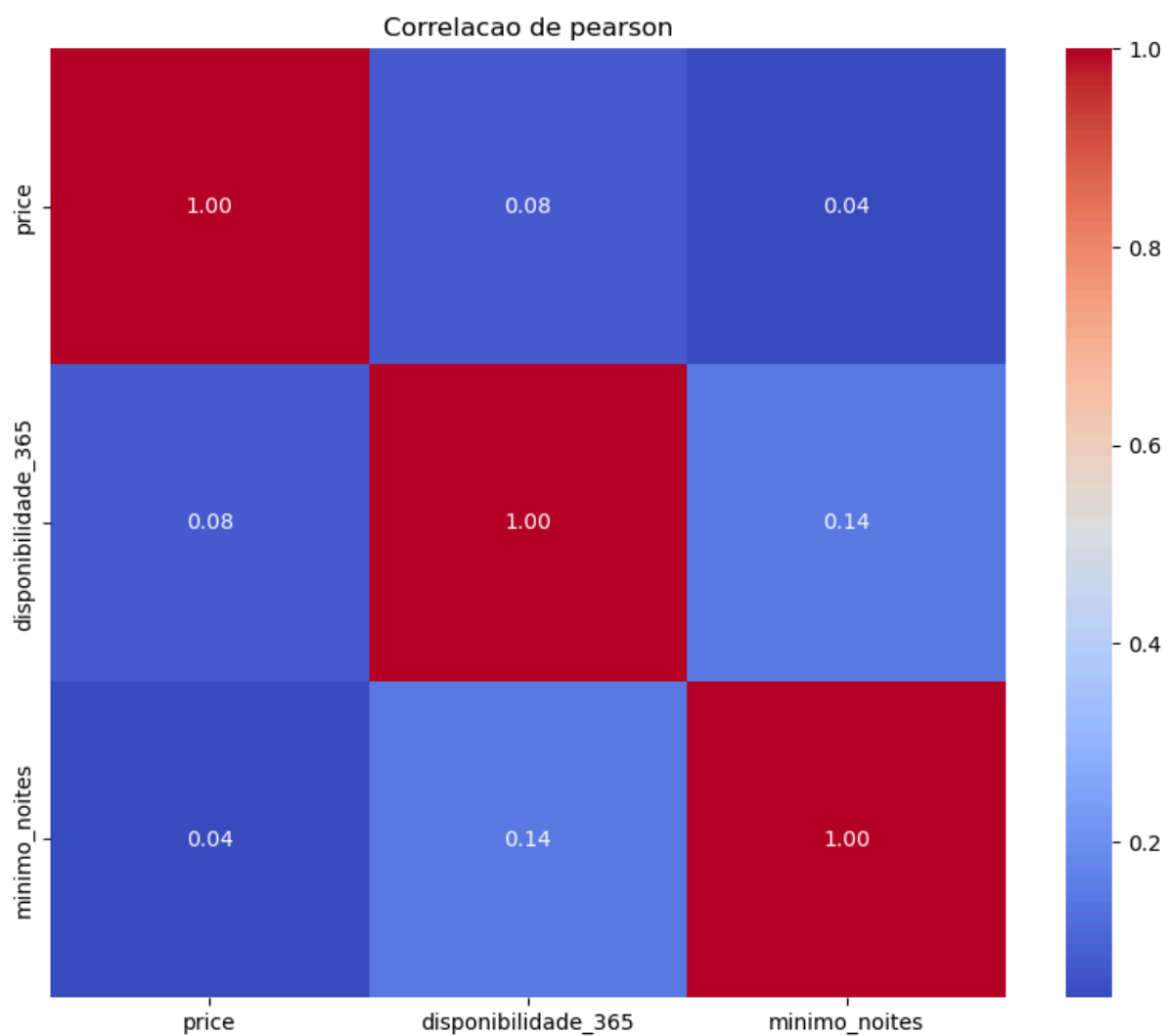
Apenas para confirmar o que os gráficos dizem, selecionei preço, o tipo do quarto e o bairro onde se encontram os locais mais baratos, conforme a figura abaixo

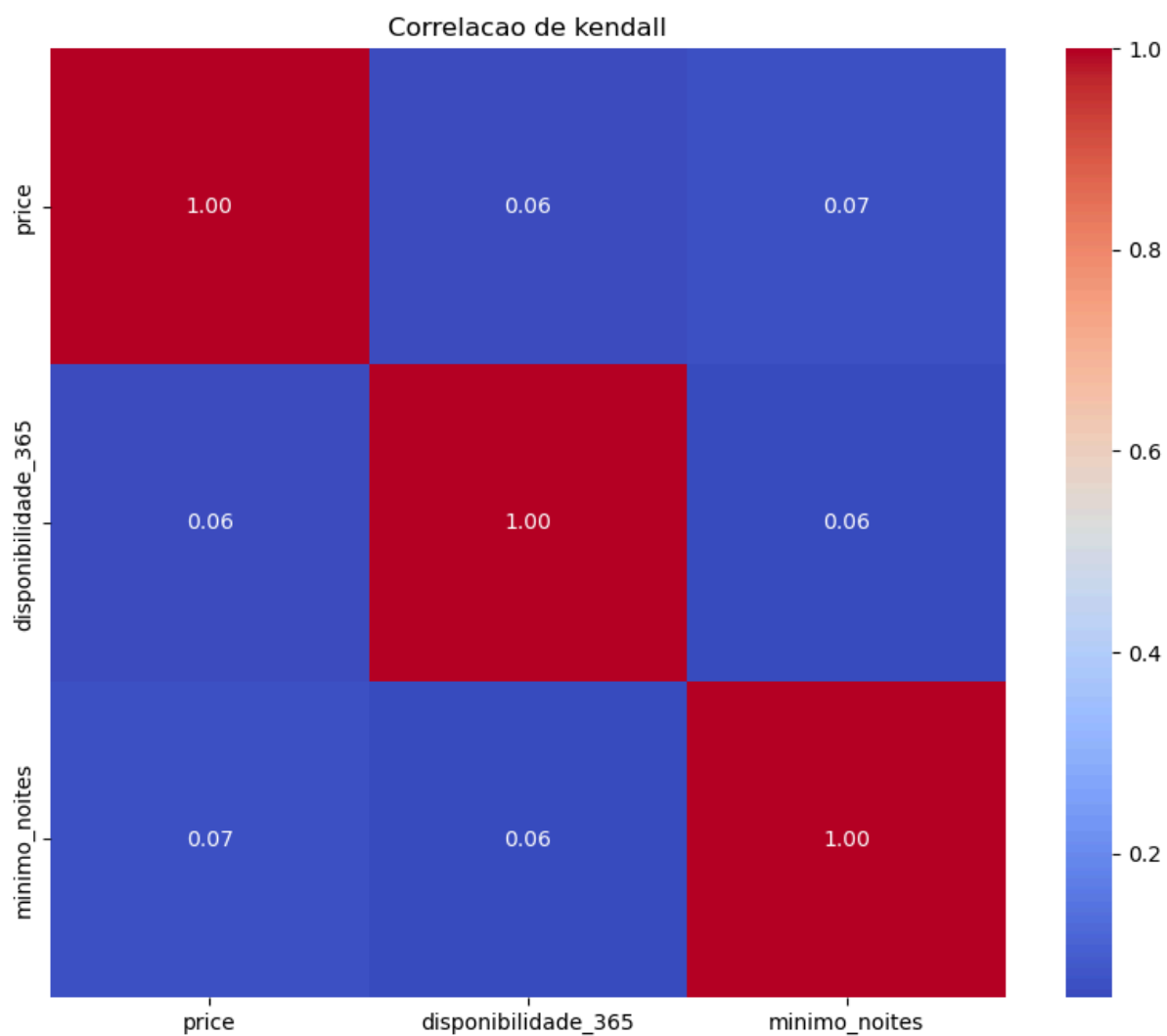
```
sort_df_rent[sort_df_rent['price'] == 10]
```

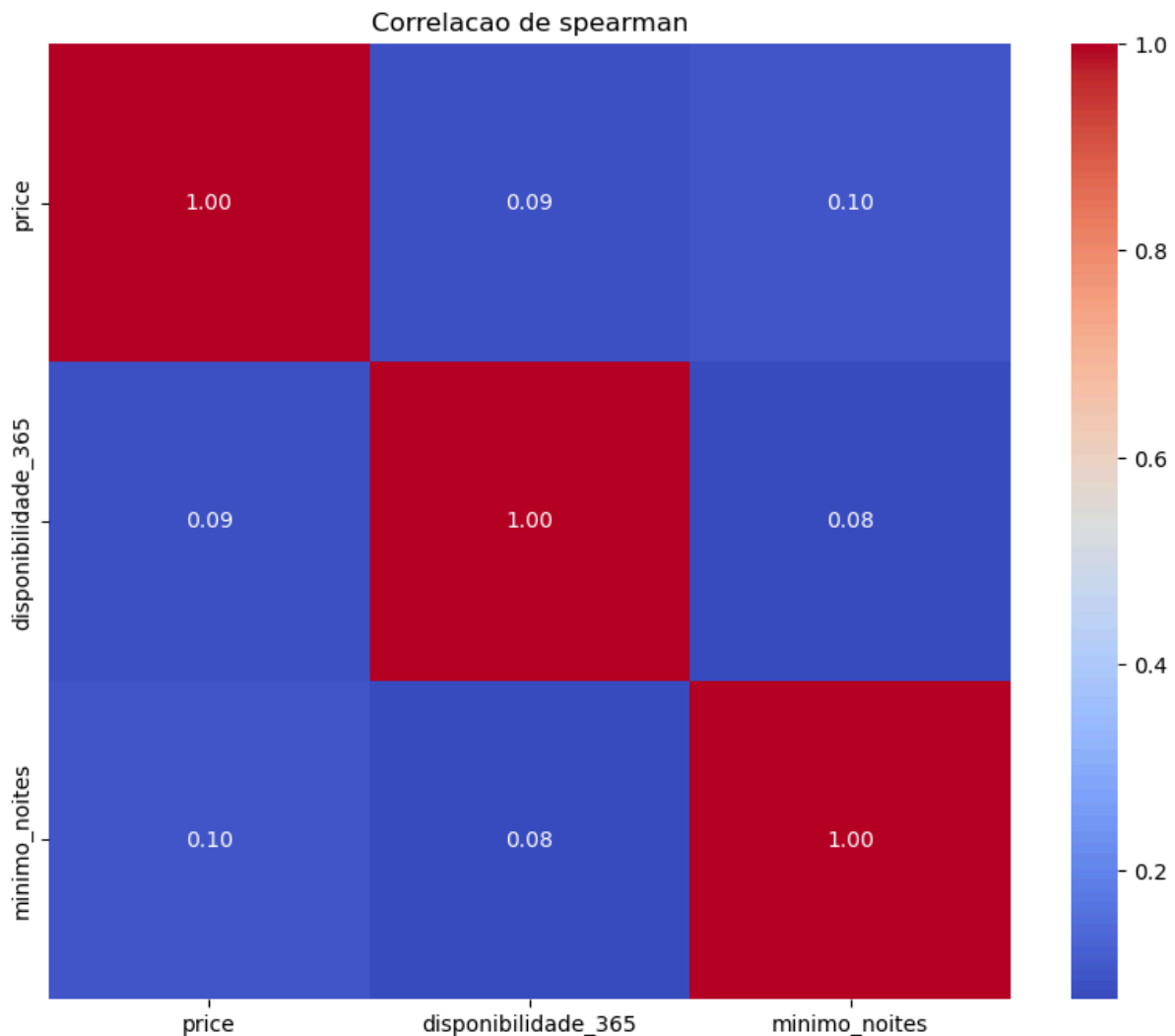
	bairro	price	room_type
35600	Bedford-Stuyvesant	10	Entire home/apt
32809	Sunset Park	10	Entire home/apt
27971	Greenpoint	10	Entire home/apt
33224	Jamaica	10	Entire home/apt
22834	Jamaica	10	Entire home/apt
23255	Upper East Side	10	Entire home/apt
2859	East Village	10	Entire home/apt

Q2: O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Para responder a esta questão, procurei ver a correlação através do coeficiente de Pearson, correlação de Kendall e a correlação de Spearman. Fiz os mapas de calor de cada para melhor visualização:





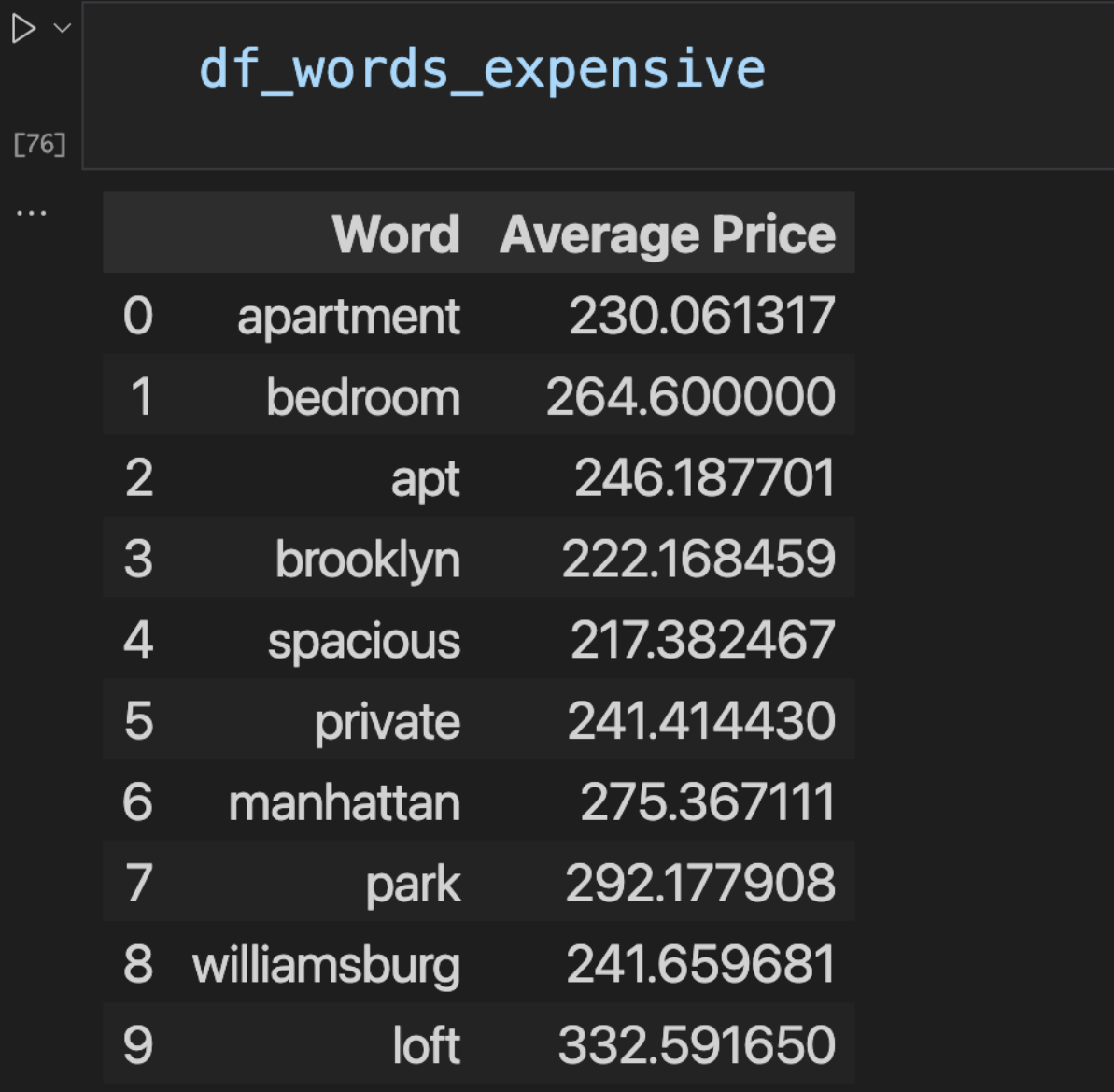


Dos gráficos, é possível inferir que o número mínimo de noites, a disponibilidade durante o ano e o preço estão relacionados, sendo a disponibilidade durante o ano e o preço terem um maior *score* na correlação de Pearson de 0.14 .

Q3: Existe algum padrão no texto do nome do local para lugares de mais alto padrão?

Inicialmente fiz uma manipulação de forma a analisar a média do preço por bairro e procurei os valores que estão acima dessa média por bairro e os separei em um outro dataframe intitulado de *df_price_abov_avg*. Em seguida, utilizei uma análise de tokens, aplicando técnicas de *NLP* para extrair as 10 palavras que mais estão presentes nos imóveis de alto padrão. Assim, as

palavras que sugerem uma relação com imóveis de alto padrão são as seguintes:



A screenshot of a Jupyter Notebook interface. At the top, a code cell contains the text `df_words_expensive`. Below it, a console cell shows the output of the code, which is a DataFrame. The DataFrame has two columns: 'Word' and 'Average Price'. It contains 10 rows of data, sorted by average price in descending order. The words listed are: apartment, bedroom, apt, brooklyn, spacious, private, manhattan, park, williamsburg, and loft. The average prices range from 230.061317 to 332.591650.

	Word	Average Price
0	apartment	230.061317
1	bedroom	264.600000
2	apt	246.187701
3	brooklyn	222.168459
4	spacious	217.382467
5	private	241.414430
6	manhattan	275.367111
7	park	292.177908
8	williamsburg	241.659681
9	loft	332.591650

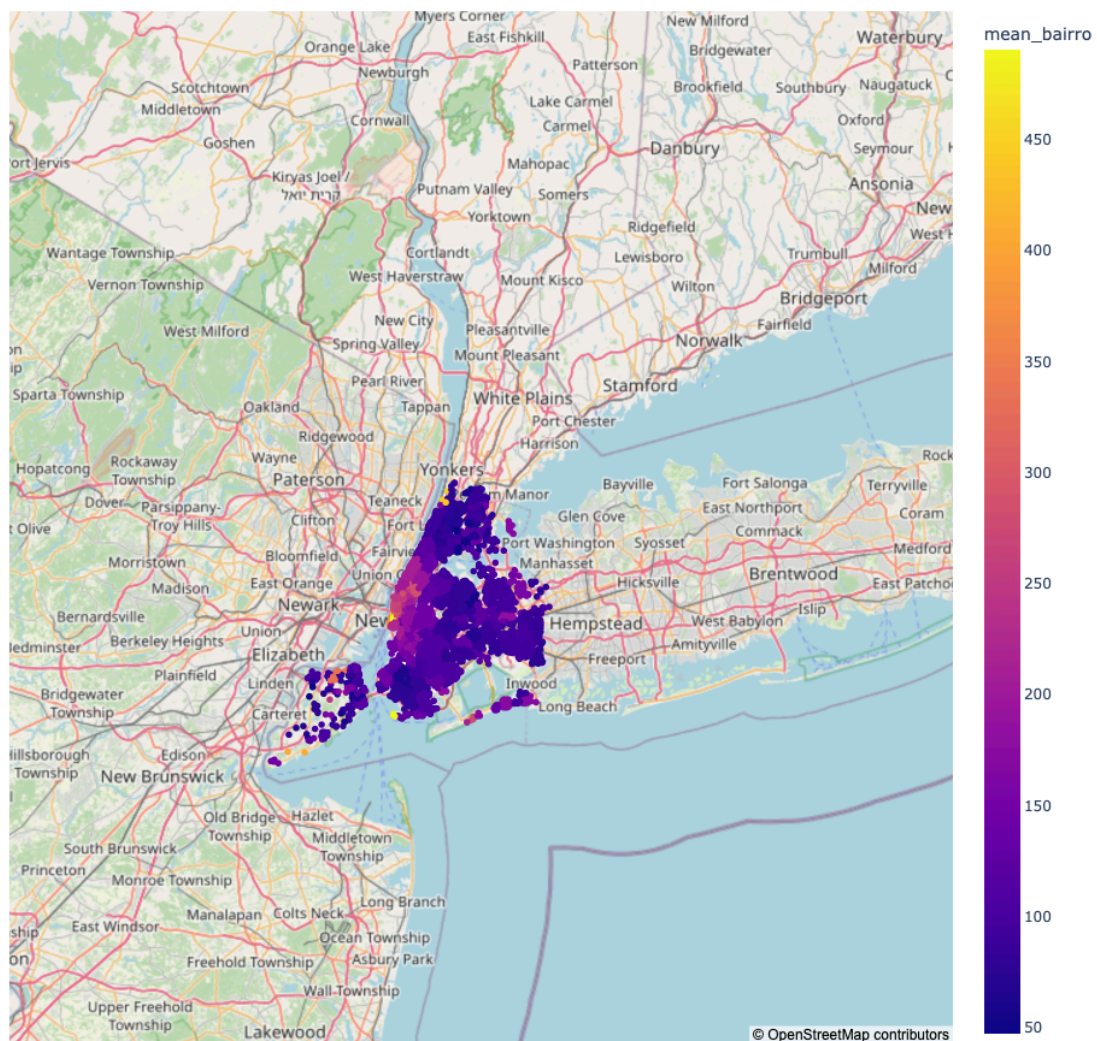
Para efeito de comparação, optei por ordenar os imóveis de alto padrão e comparar com as 10 palavras que mais estão presentes e perceber que todas

estão presente no dataframe.

```
df_most_expensive = df_price_abov_avg.sort_values(by = 'price', ascending=False)
df_most_expensive.head(15)
```

	id	nome	bairro	mean_bairro	bairro_group	price
17691	13894339	Luxury 1 bedroom apt. -stunning Manhattan views	Greenpoint	144.95	Brooklyn	10000
9150	7003697	Furnished room in Astoria apartment	Astoria	117.19	Queens	10000
29226	22436899	1-BR Lincoln Center	Upper West Side	210.92	Manhattan	10000
6529	4737930	Spanish Harlem Apt	East Harlem	133.20	Manhattan	9999
40421	31340283	2br - The Heart of NYC: Manhattans Lower East ...	Lower East Side	186.31	Manhattan	9999
12341	9528920	Quiet, Clean, Lit @ LES & Chinatown	Lower East Side	186.31	Manhattan	9999
30256	23377410	Beautiful/Spacious 1 bed luxury flat-TriBeCa/Soho	Tribeca	490.64	Manhattan	8500
4376	2953058	Film Location	Clinton Hill	181.89	Brooklyn	8000
29650	22779726	East 72nd Townhouse by (Hidden by Airbnb)	Upper East Side	188.95	Manhattan	7703
45654	34895693	Gem of east Flatbush	East Flatbush	104.22	Brooklyn	7500
42511	33007610	70' Luxury MotorYacht on the Hudson	Battery Park City	367.56	Manhattan	7500
44022	33998396	3000 sq ft daylight photo studio	Chelsea	249.74	Manhattan	6800
37182	29547314	Apartment New York \nHell's Kitchens	Upper West Side	210.92	Manhattan	6500
48031	36056808	Luxury TriBeCa Apartment at an amazing price	Tribeca	490.64	Manhattan	6500
3773	2271504	SUPER BOWL Brooklyn Duplex Apt!!	Clinton Hill	181.89	Brooklyn	6500

Para efeito de visualização optei por fazer o *plot* iterativo da média por bairros (A interatividade é possível ao abrir o Notebook):



Exercício de predição

Nesta seção diferentes técnicas de regressão foram utilizadas, uma vez que estamos procurando predizer um valor. Para isso foi tirado qualquer ruído do restante no dataset. Uma das transformações necessárias foi identificar cada bairro(do tipo *object*) para um tipo inteiro(*int*) usando a função *LabelEncoded* para ser possível colocá-la na variável X, ou seja, as *features* as colunas *id*, *host_id*, *mean_bairro*, *bairro_encoded*, *latitude*, *longitude*, *minimo_noites*, *numero_de_reviews*, *reviews_por_mes*, *calculado_host_listings_count*, *disponibilidade_365*. Basicamente as colunas com dados numéricos. Optei primeiramente por uma abordagem mais clássica usando o *LinearRegressor()* e o *fit()* porém o modelo estava

generalizando pouco, cerca de 9% apenas. Utilizei por fim o modelo *RandomForest* e o *Cross-Validation* para obter resultados melhores. Utilizei a métrica *Mean Absolute Error(MAE)* e essa métrica mede a diferença em módulo entre os valores que foram preditos e os valores reais para avaliar o modelo. Obtive uma MAE de 24.5 e uma vez que o range de *X_scaled* varia de 10 e 10.000 pode-se dizer agora que o modelo generaliza muito bem.