Lighthouse: Desafio de Ciência de Dados

Lara Esquivel de Brito Santos

Primeiros Passos

- Compreender o dicionário de Dados
- Os tipos de Dados
- Verifiquei quais são vazios
- Estatísticas Rápidas

Estatísticas Rápidas: Numéricas

	id	host_id	latitude	longitude	price	minimo_noites	numero_de_reviews	reviews_por_mes	calculado_host_listings_count	disponibilidade_365
count	4.889400e+04	4.889400e+04	48894.000000	48894.000000	48894.000000	48894.000000	48894.000000	38842.000000	48894.000000	48894.000000
mean	1.901753e+07	6.762139e+07	40.728951	-73.952169	152.720763	7.030085	23.274758	1.373251	7.144005	112.776169
std	1.098288e+07	7.861118e+07	0.054529	0.046157	240.156625	20.510741	44.550991	1.680453	32.952855	131.618692
min	2.595000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.472371e+06	7.822737e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967743e+07	3.079553e+07	40.723075	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915225e+07	1.074344e+08	40.763117	-73.936273	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

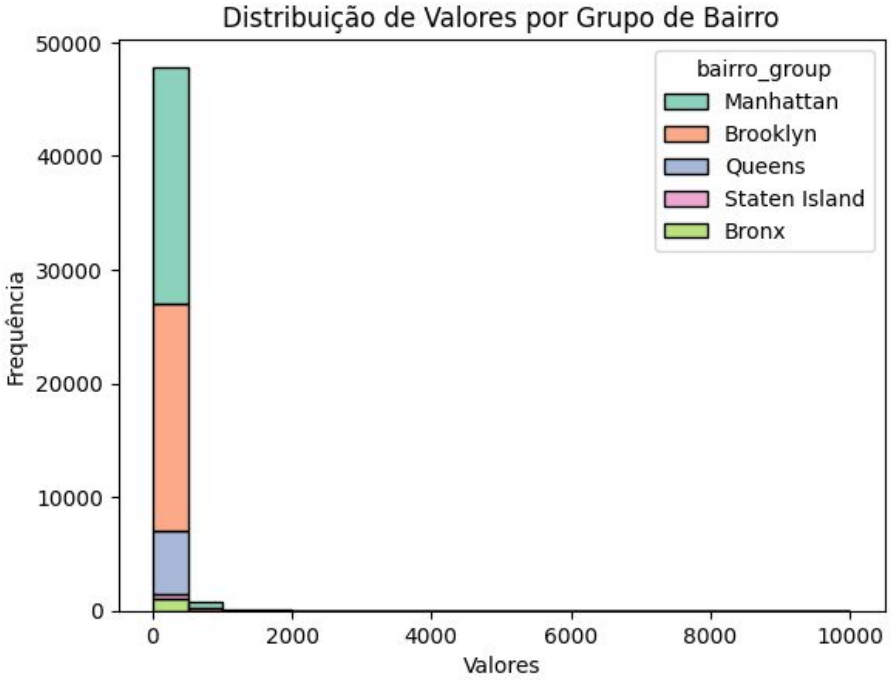
Figura 1: Estatística gerada pelo Dataframe.describe()

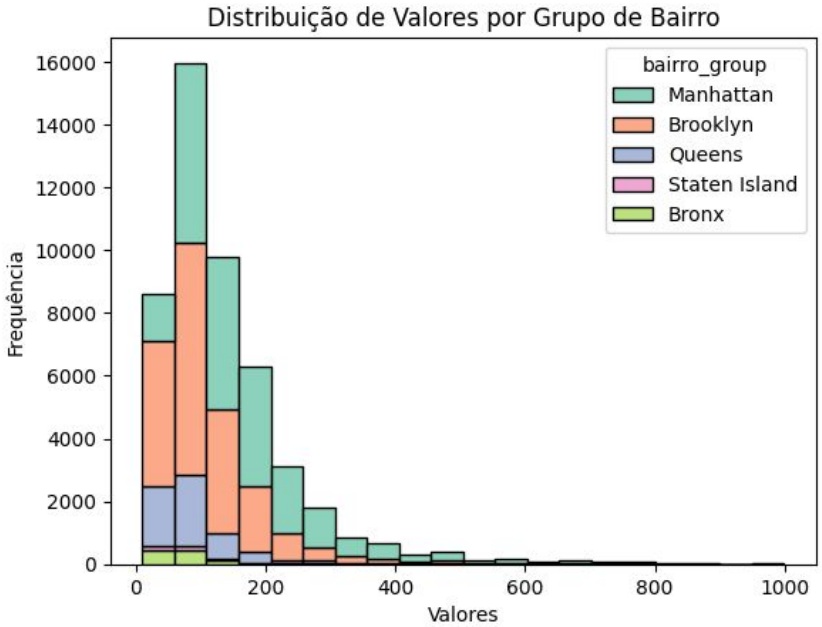
Estatísticas Rápidas: Categóricas

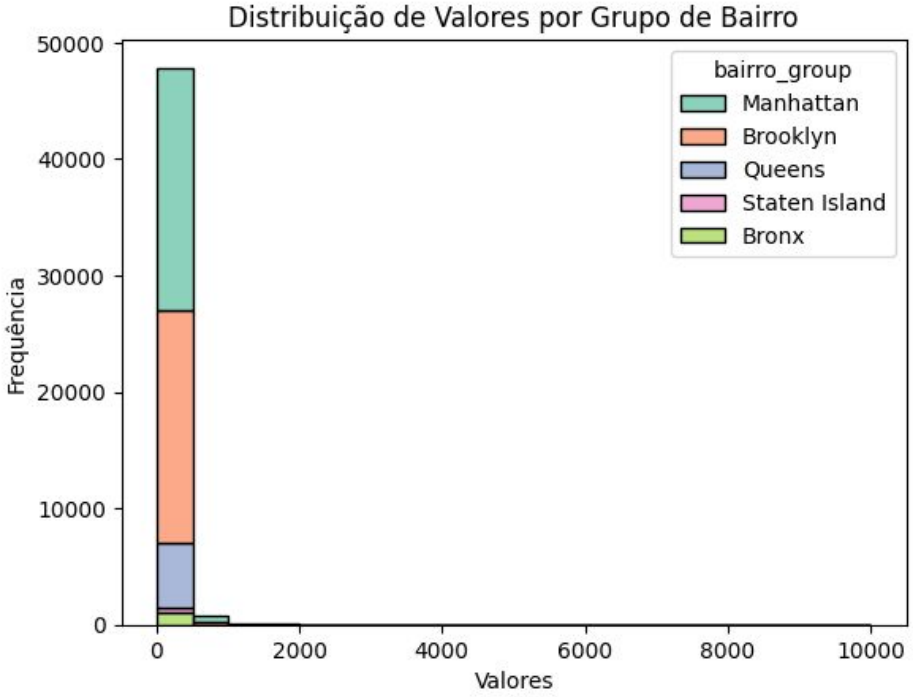
	nome	host_name	bairro_group	bairro	room_type	ultima_review
count	48878	48873	48894	48894	48894	38842
unique	47904	11452	5	221	3	1764
top	Hillside Hotel	Michael	Manhattan	Williamsburg	Entire home/apt	2019-06-23
freq	18	417	21661	3920	25409	1413

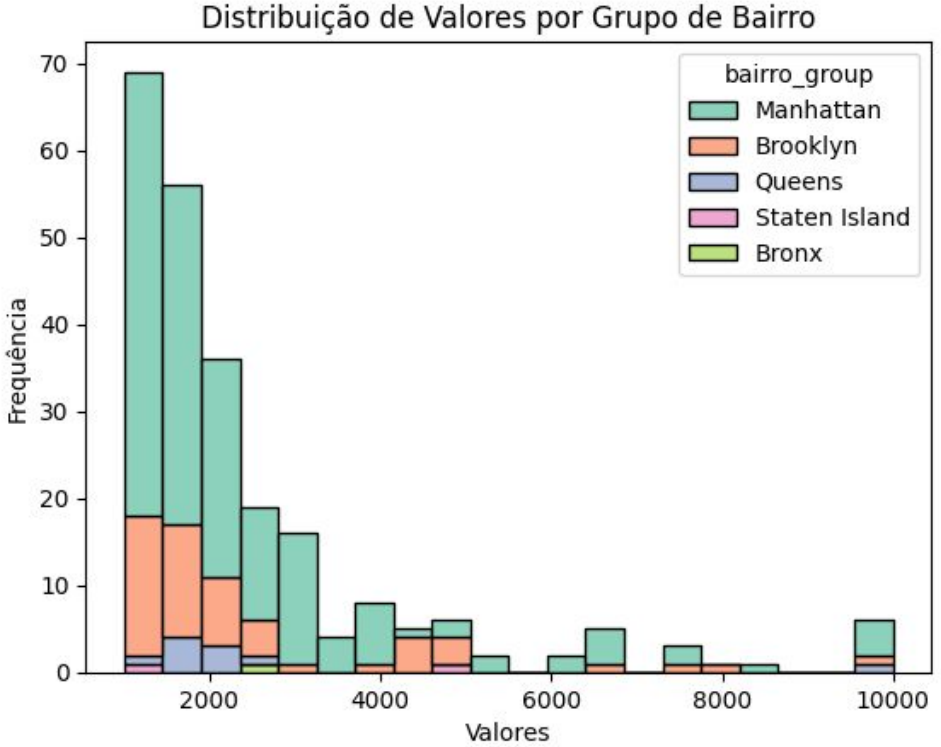
Figura 2: Estatística gerada pelo Dataframe.describe()

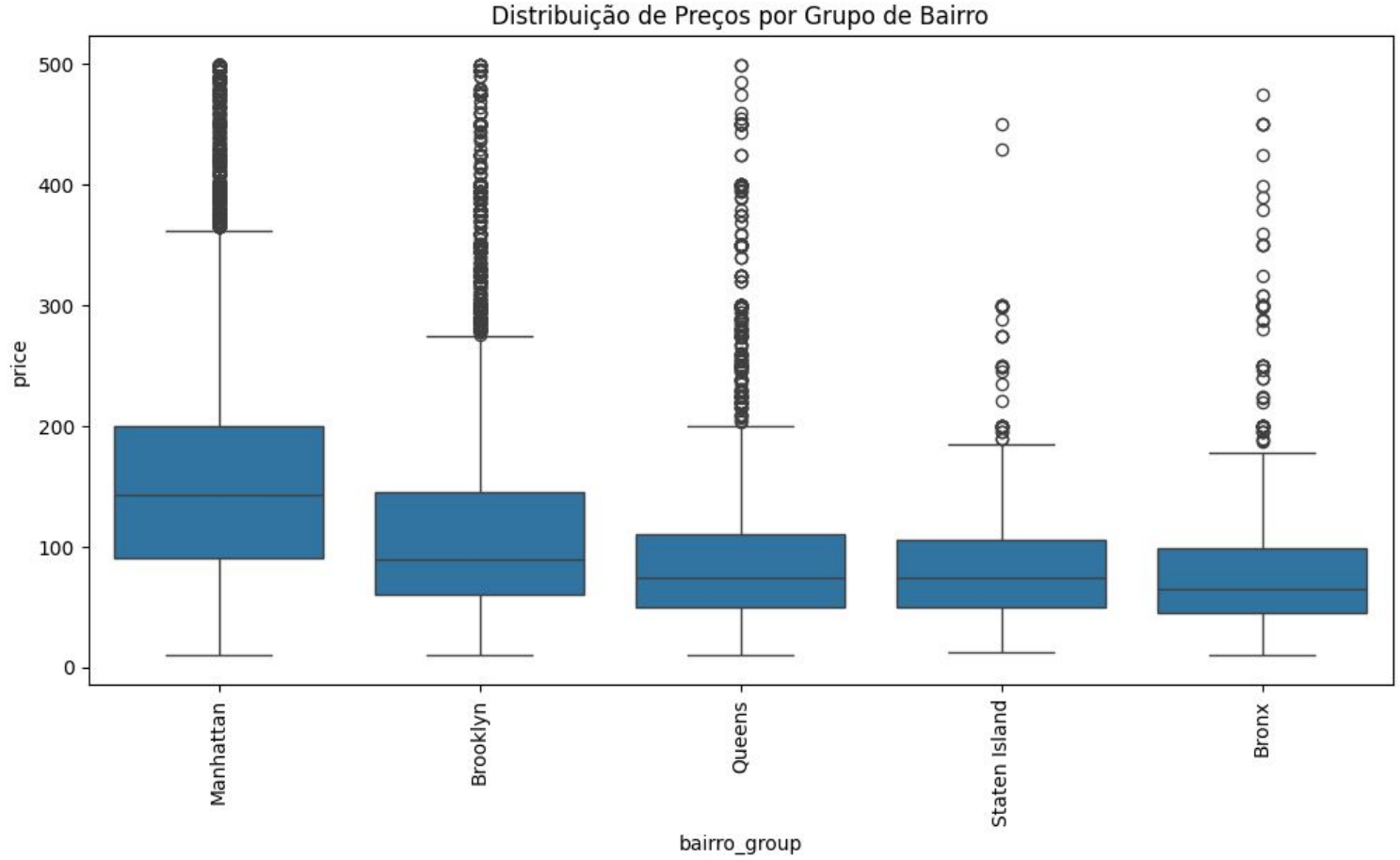
Distribuição

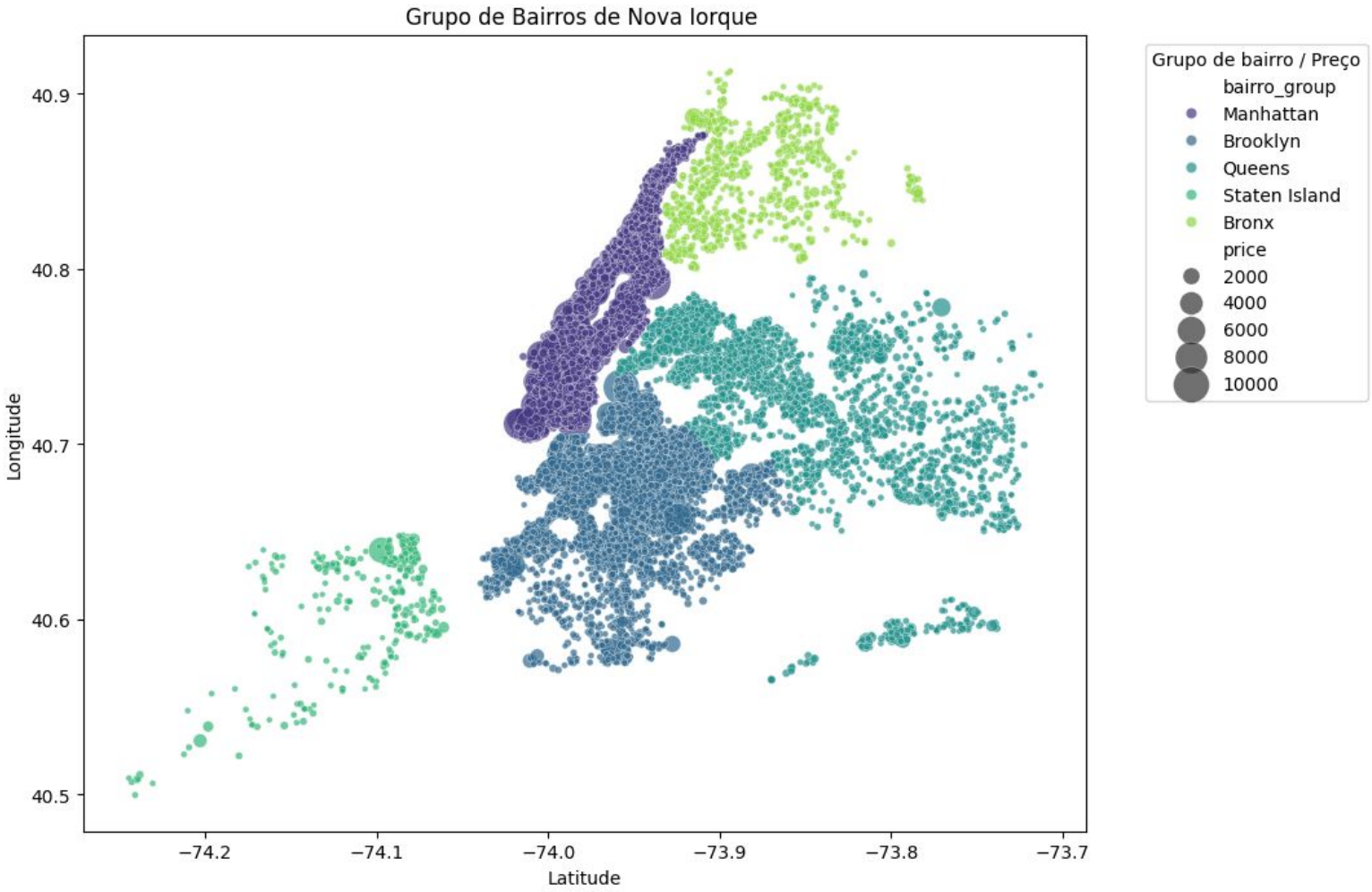





Distribuição de Valores por Grupo de Bairro
bairro_group
Manhattan
Brooklyn
Queens
Staten Island
Bronx
Frequência
Valores
0
10000
20000
30000
40000
50000
0
2000
4000
6000
8000
10000

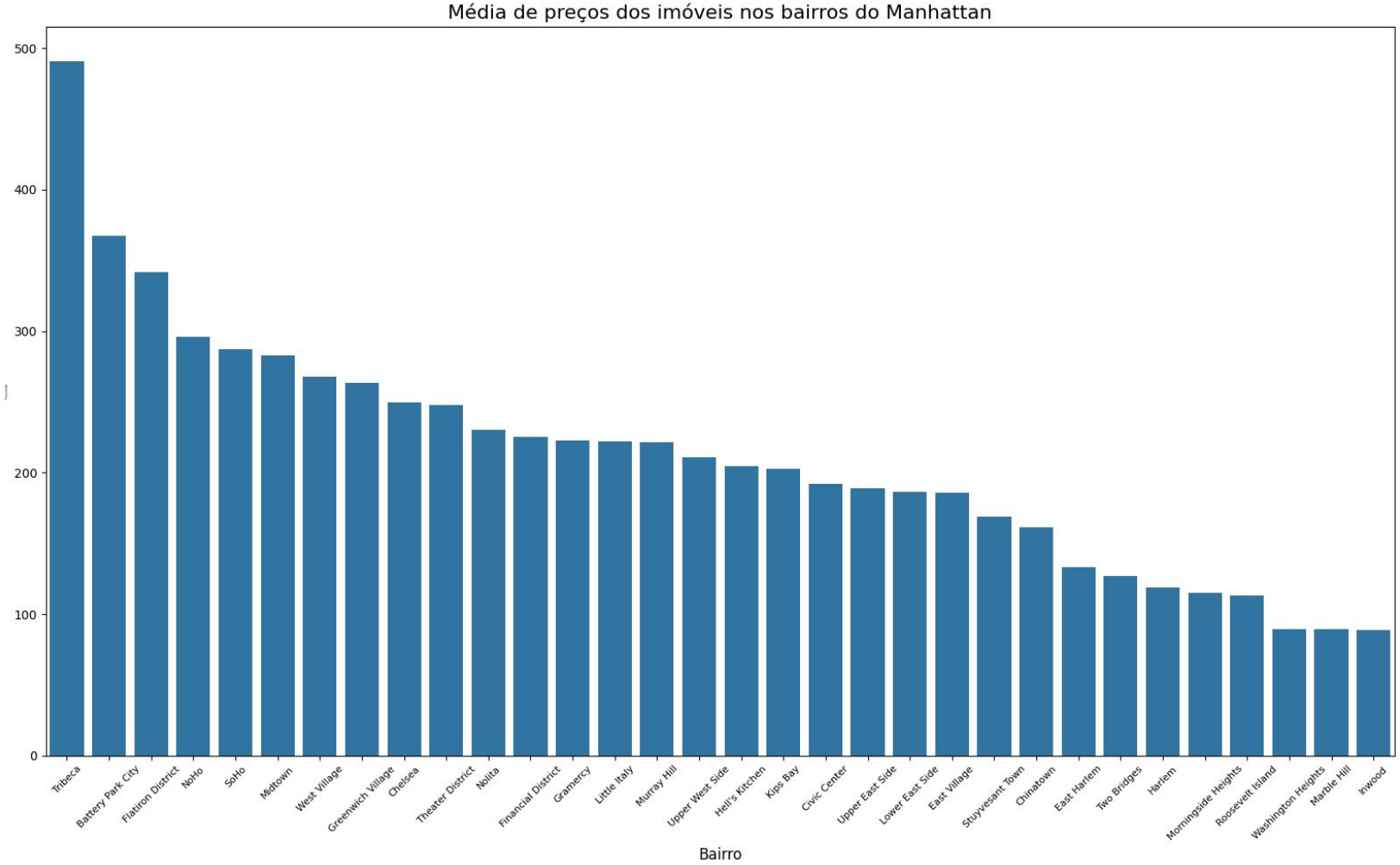

Distribuição de Valores por Grupo de Bairro
bairro_group
Manhattan
Brooklyn
Queens
Staten Island
Bronx
Frequência
Valores
0
10
20
30
40
50
60
70
2000
4000
6000
8000
10000


Distribuição de Preços por Grupo de Bairro
price
500
400
300
200
100
0
Manhattan
Brooklyn
Queens
Staten Island
Bronx
bairro_group

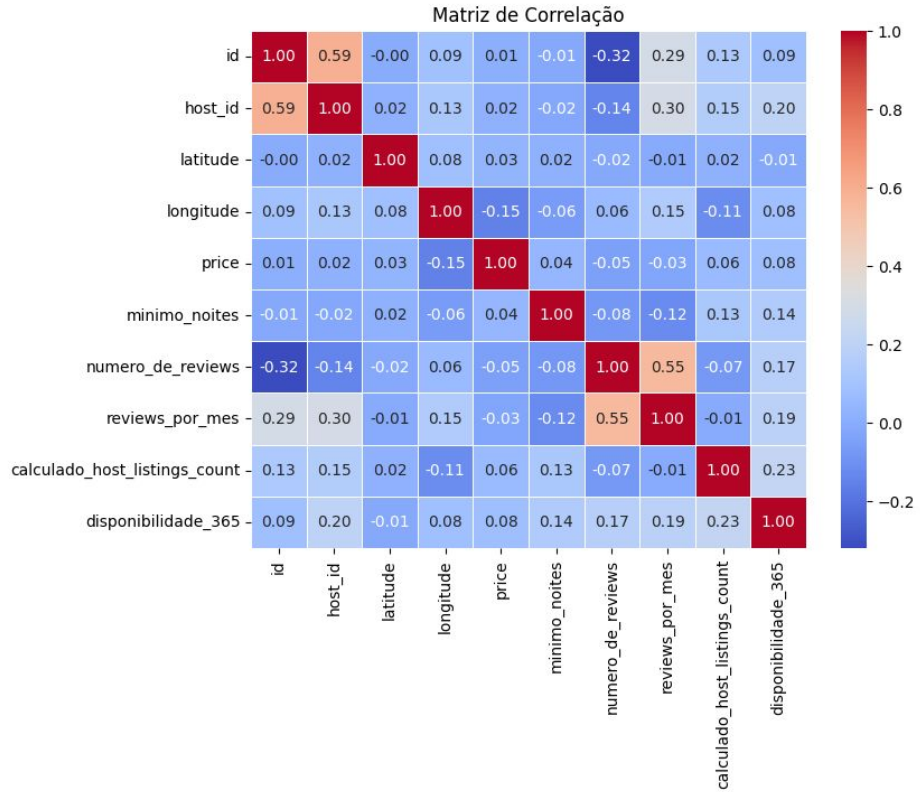

Grupo de Bairros de Nova Iorque
Longitude
Latitude
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Grupo de bairro / Preço
bairro_group
Manhattan
Brooklyn
Queens
Staten Island
Bronx
price
2000
4000
6000
8000
10000

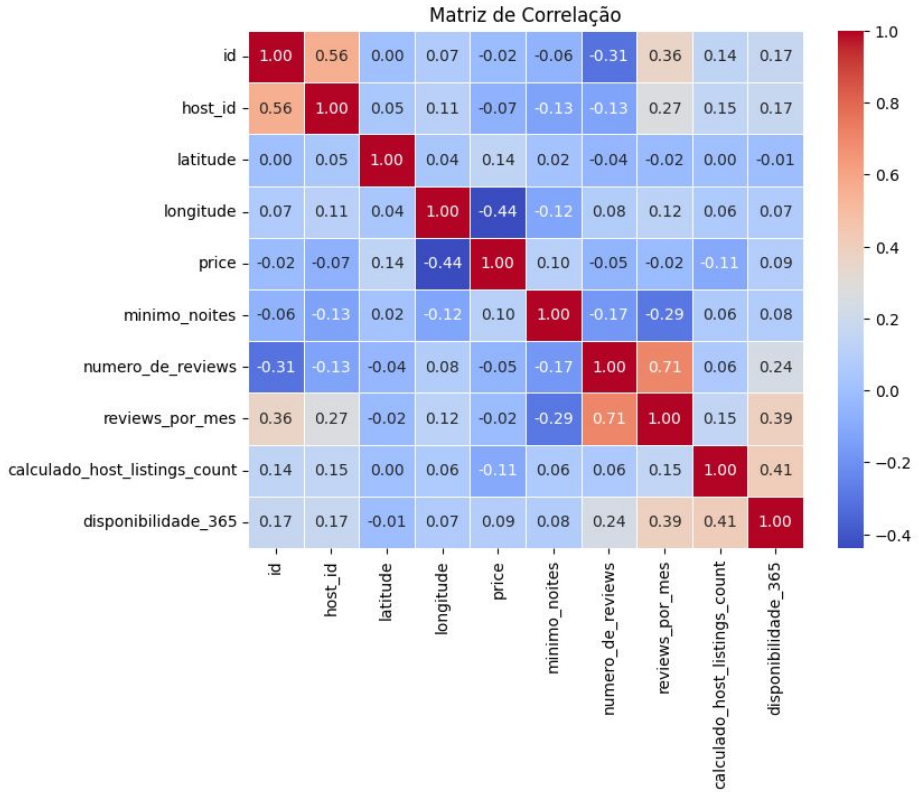
Onde comprar um imóvel?

- Ao gerar a tabela de atributos categóricos do grupo de bairro Staten Island, o nome de anúncio mais frequente é “New York home ferry from Manhattan”
- Comecei a buscar a palavra Manhattan entre os anúncios dos outros bairros
- Revi as tabelas de descrição categóricas, as de Manhattan, o bairro com o maior número de anúncios é o Harlem que não possui uma média de valor tão alta.


Média de preços dos imóveis nos bairros do Manhattan
500
400
300
200
100
0
Tribeca
Battery Park City
Flatiron District
NoHo
SoHo
Midtown
West Village
Greenwich Village
Chelsea
Theater District
Nolita
Financial District
Gramercy
Little Italy
Murray Hill
Upper West Side
Hell's Kitchen
Kips Bay
Civic Center
Upper East Side
Lower East Side
East Village
Stuyvesant Town
Chinatown
East Harlem
Two Bridges
Harlem
Morningside Heights
Roosevelt Island
Washington Heights
Marble Hill
Inwood
Bairro

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





Teste de Hipótese

```
Qui-quadrado: 2254537.563705246
Valor p: 0.0
Graus de liberdade: 2071656
Frequências esperadas:
[[2.08105067e+00 1.91363051e+00 1.30892130e+00 ... 4.90968230e-04
  1.63656077e-04 1.63656077e-04]
 [2.60131334e-01 2.39203813e-01 1.63615163e-01 ... 6.13710288e-05
  2.04570096e-05 2.04570096e-05]
 [2.60131334e-01 2.39203813e-01 1.63615163e-01 ... 6.13710288e-05
  2.04570096e-05 2.04570096e-05]
 ...
 [2.60131334e-01 2.39203813e-01 1.63615163e-01 ... 6.13710288e-05
  2.04570096e-05 2.04570096e-05]
 [5.20262668e-01 4.78407626e-01 3.27230325e-01 ... 1.22742058e-04
  4.09140192e-05 4.09140192e-05]
 [2.60131334e-01 2.39203813e-01 1.63615163e-01 ... 6.13710288e-05
  2.04570096e-05 2.04570096e-05]]
Rejeita-se a hipótese nula: As variáveis não são independentes.
```

Existe padrão nos nomes das localidades dos imóveis de alto padrão?

- Definir imóvel de alto padrão acima de $1000,00
- Maior parte dos imóveis estão em Manhattan, 172 no total, seguido do Brooklyn 54 e depois o Queens com 10
- Verifiquei os bairros de cada grupo, metade dos bairros de Manhattan possui imóveis nessa faixa.
- Verifiquei usando a biblioteca spacy por localidades e encontrei Manhattan, West Village, East Village, Williamsburg, Midtown, nomes de bairros nobres de Manhattan e a presença da palavra Superbowl.

O que estamos buscando?

- O preço de aluguéis, logo é um modelo regressivo
- Só não foram utilizados o nome, os identificadores, nome do host e a ultima_review.
- Bairro e grupo de bairro foram representados por médias de preço
- Para tipo de quarto foi utilizado LabelEnconder
- Algoritmo escolhido foi o RandomForest e a métrica Coeficiente de Determinação
- Valor previsto foi de 213,185