Indicium
Desafio Cientista de Dados
Relatórios das análises estatísticas e EDA

1. Estatísticas

Comandos do pandas.py utilizados para visualizacao de algumas características do arquivo .csv utilizado:

```python
data.head()
data.columns
data.count()
data.shape
data.isna().sum()
data.describe()
data['price'].median()
data['price'].mean()
data['bairro_group'].value_counts()
data.info()
```

data.head()	
data.columns	Index(['id', 'nome', 'host_id', 'host_name', 'bairro_group', 'bairro', 'latitude', 'longitude', 'room_type', 'price', 'minimo_noites', 'numero_de_reviews', 'ultima_review', 'reviews_por_mes', 'calculado_host_listings_count', 'disponibilidade_365'], dtype='object')
data.count()	id: 48894 nome: 48878 host_id: 48894 host_name: 48873 bairro_group: 48894 bairro: 48894 latitude: 48894 longitude: 48894 room_type: 48894 price: 48894 minimo_noites: 48894 numero_de_reviews: 48894 ultima_review: 38842 reviews_por_mes: 38842 calculado_host_listings_count: 48894 disponibilidade_365: 48894

	dtype: int64
data.shape()	(48894, 16)
data.isna().sum()	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
data.describe()	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

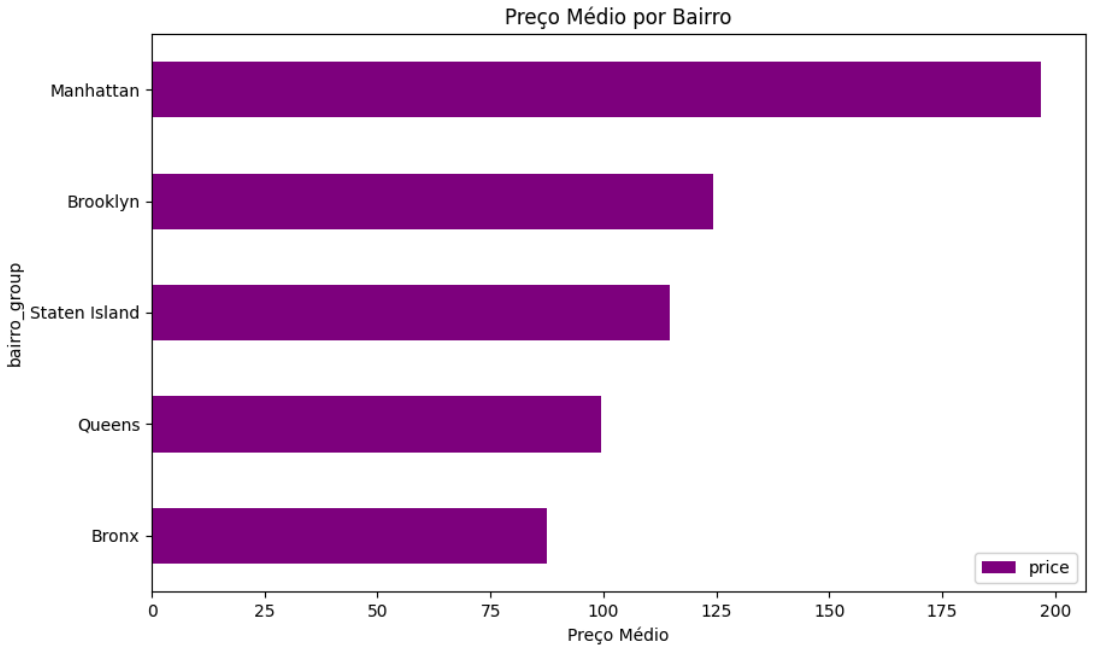
	75% 2.915225e+07 1.074344e+08 40.763117 -73.936273 175.000000 5.000000 24.000000 2.020000 2.000000 227.000000 max 3.648724e+07 2.743213e+08 40.913060 -73.712990 10000.000000 1250.000000 629.000000 58.500000 327.000000 365.000000
data[‘price’].median()	106.0
data[‘price’].mean()	152.7207632838385
data[‘bairro_group’].value_counts()	bairro_group Manhattan: 21661 Brooklyn : 20103 Queens: 5666 Bronx: 1091 Staten Island: 373 Name: count, dtype: int64
data.info()	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
data[‘price’].max()	10000
data[‘price’].min()	0

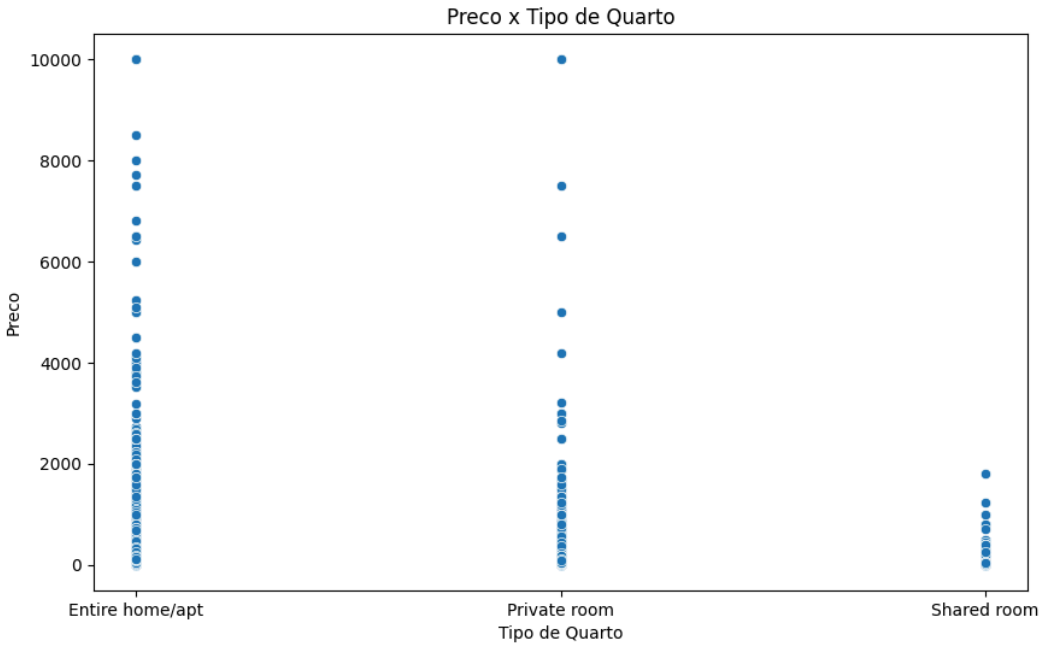
2. Tratamento dos dados

```python
#dados faltantes de strings
data['nome'] = data['nome'].fillna('Sem informacao')
data['bairro_group'] = data['bairro_group'].fillna('Sem informacao do grupo do bairro')
data['bairro'] = data['bairro'].fillna('Sem informacao do bairro')
data['room_type'] = data['room_type'].fillna('Sem informacao do tipo de quarto')
data['host_name'] = data['host_name'].fillna('Sem informaccoes do anfitriao')

#dados faltantes de númericos
data['numero_de_reviews'] = data['numero_de_reviews'].fillna(0)
data['reviews_por_mes'] = data['reviews_por_mes'].fillna(0)
data['disponibilidade_365'] = data['disponibilidade_365'].fillna(0)

#valores com datas
data['ultima_review'] = data['ultima_review'].fillna('1800-06-07')
```

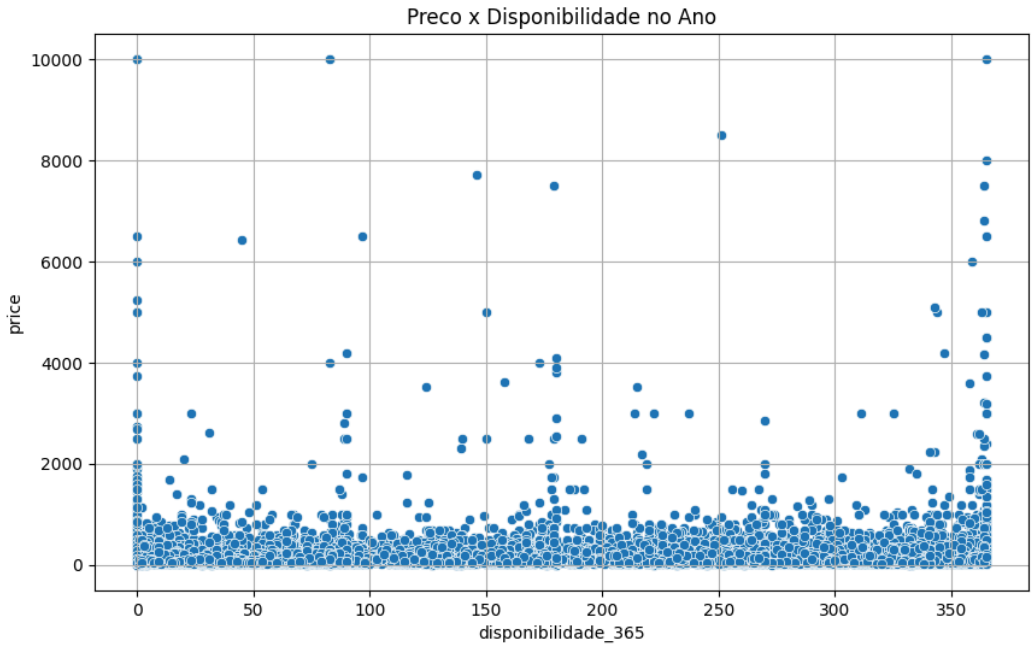
3. Gráficos com alguns resultados
   3.1 - Preço Médio por Bairros



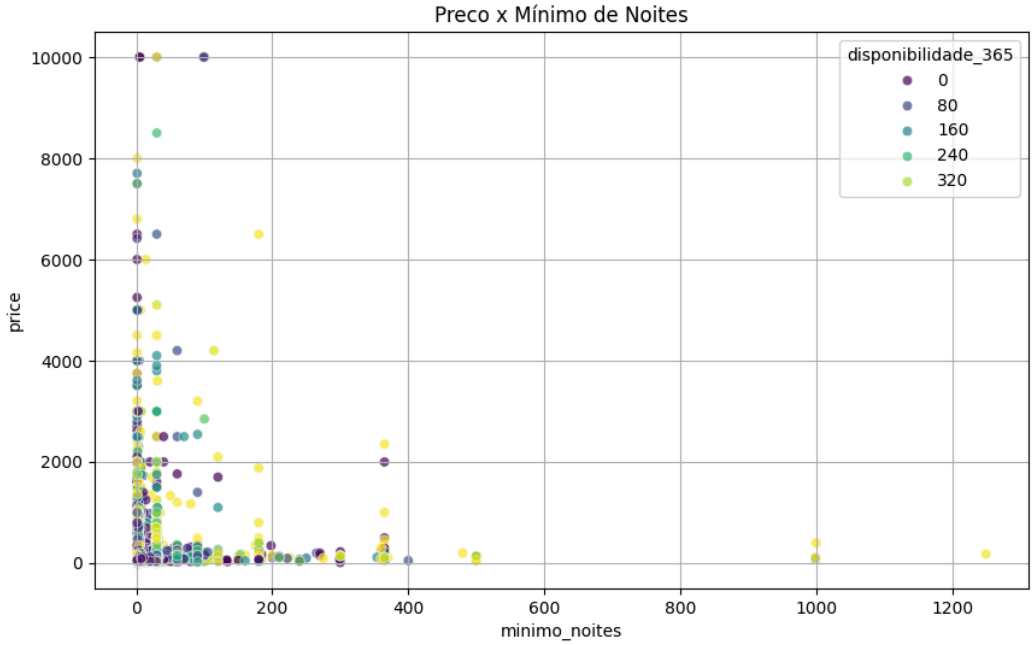
3.2 - Preço por tipo de Quarto



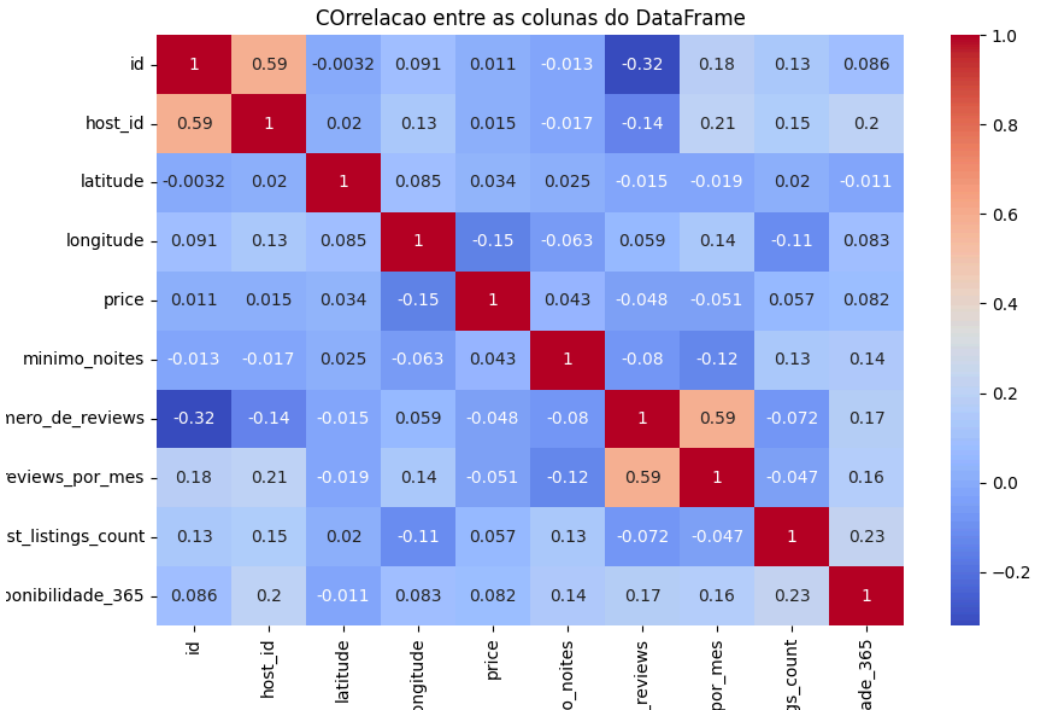
3.3 - Preço e disponibilidade



3.4 - Preço x mínimo de noites



3.5 - Correlação entre os valores numéricos do DataFrame


COrrelacao entre as colunas do DataFrame
id 1 0.59 -0.0032 0.091 0.011 -0.013 -0.32 0.18 0.13 0.086
host_id 0.59 1 0.02 0.13 0.015 -0.017 -0.14 0.21 0.15 0.2
latitude -0.0032 0.02 1 0.085 0.034 0.025 -0.015 -0.019 0.02 -0.011
longitude 0.091 0.13 0.085 1 -0.15 -0.063 0.059 0.14 -0.11 0.083
price 0.011 0.015 0.034 -0.15 1 0.043 -0.048 -0.051 0.057 0.082
minimo_noites -0.013 -0.017 0.025 -0.063 0.043 1 -0.08 -0.12 0.13 0.14
nero_de_reviews -0.32 -0.14 -0.015 0.059 -0.048 -0.08 1 0.59 -0.072 0.17
eviews_por_mes 0.18 0.21 -0.019 0.14 -0.051 -0.12 0.59 1 -0.047 0.16
st_listings_count 0.13 0.15 0.02 -0.11 0.057 0.13 -0.072 -0.047 1 0.23
onibilidade_365 0.086 0.2 -0.011 0.083 0.082 0.14 0.17 0.16 0.23 1
id
host_id
latitude
ongitude
price
o_noites
_reviews
por_mes
gs_count
ade_365
1.0
0.8
0.6
0.4
0.2
0.0
−0.2