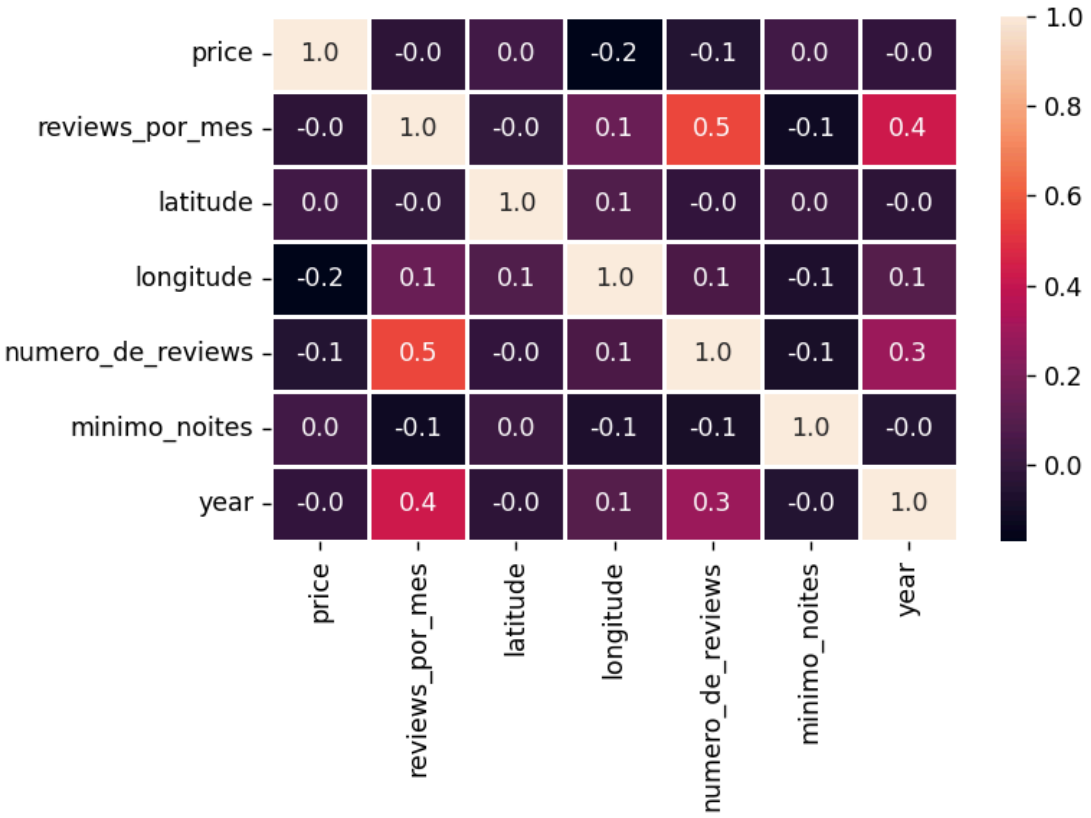
Desafio Cientista de Dados

1. **Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas.**

A fim de compreender melhor o conjunto de dados e a relação entre as possíveis variáveis, foi aplicada uma matriz de correlação. Esta matriz não trouxe um resultado de correlação satisfatório quanto ao principal objetivo das análises: entender a precificação e a tendência de mercado. Entretanto, para obter mais informações, utilizei variáveis comuns no mercado imobiliário, como localização e tipo de acomodação, para entender melhor o comportamento dos dados, buscando padrões e tendências que pudessem oferecer insights relevantes.

Entretanto, durante o processo, foi possível perceber a existência de datas de review faltando. Datas que estavam sendo usadas para poder fazer previsões futuras baseadas nos anos. Para corrigir o problema, foi criada uma coluna com o nome 'year', que armazenava o ano correspondente à data, e, para preencher as datas faltantes, foi utilizado o cálculo da média dos anos disponíveis.



Para entender a relação das variáveis, utilizei uma função que calcula o preço mínimo, médio e máximo das acomodações, baseadas nas áreas em que estão localizadas (Manhattan, Brooklyn, Bronx, Staten Island e Queens) e no tipo de acomodação ('Shared room', 'Private Room', 'Entire apartment/home'). Com o resultado, foi possível determinar que bairros como Manhattan e Brooklyn foram as áreas em que as acomodações têm um preço mais elevado.

Além disso, foi possível determinar que acomodações do tipo **‘Entire apartment/home’** têm um valor maior dentre as demais, apesar de não ser um fator mandatório, já que, para a grande maioria dos bairros e imóveis, os valores mínimos não passaram de $10 a diária.

Assim, para entender melhor a variação de preço, olhei para a descrição dos anúncios para entender a variação, ou falta dela, nos preços. Ficou claro que anúncios com descrições mais luxuosas ou de locais famosos, como **“Luxury apartment”**, **“Yacht”**, **“Mansion”**, **“House”**, **“Hells Kitchen”**, lideraram a máxima dos preços. Já anúncios com descrições mais comuns tiveram maior incidência em anúncios de valores menores.

```
----------Group: Manhattan Room Type: Entire home/apt----------

Max Price:
                                      nome bairro_group          bairro        room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
29661  East 72nd Townhouse by (Hidden by Airbnb)    Manhattan  Upper East Side  Entire home/apt              1                  0  7703.0                  146

Min Price:
                                        nome bairro_group          bairro        room_type  minimo_noites  numero_de_reviews  price  disponibilidade_365
2859   Large furnished 2 bedrooms- - 30 days Minimum    Manhattan     East Village  Entire home/apt             30                  0   10.0                  137
23255         Quiet, Cozy UES Studio Near the Subway    Manhattan  Upper East Side  Entire home/apt              3                 10   10.0                    0

Mean Price:  246.40
```

```
----------Group: Manhattan Room Type: Private room----------

Max Price:
                                  nome bairro_group          bairro     room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
37193  Apartment New York \nHell’s Kitchens    Manhattan  Upper West Side  Private room             30                  0  6500.0                   97

Min Price:
                                        nome bairro_group          bairro     room_type  minimo_noites  numero_de_reviews  price  disponibilidade_365
22286                                  Jen Apt    Manhattan            SoHo  Private room              5                  2   10.0                    0
31065  Very Spacious bedroom, steps from CENTRAL PARK.    Manhattan  Upper West Side  Private room              1                  2   10.0                    0
31406                        Cozy feel at home studio    Manhattan        Kips Bay  Private room              5                 42   10.0                    2

Mean Price:  115.54
```

Entretanto, vale ressaltar que Brooklyn e Manhattan representam juntos mais de 40 mil dos imóveis disponíveis, sendo Manhattan com 21661 imóveis, Brooklyn com 20103, Staten Island com 314, Bronx com 1091 e Queens com 5666.

Adiante, durante a análise dos preços por localização encontrei uma falta de padrão para valores acima de 8000, considerando que tanto a descrição quanto o tipo de acomodação não seguiam os padrões para acomodações de alto valor. Para corrigir esse erro utilizei a mediana dos valores encontrados por localização e tipo de acomodação.

O número de reviews e disponibilidade anual também não seguiu um padrão para valores altos ou baixos, porém ainda podem ser usadaos para algumas analises, como lugares mais frequentados pelo público.

2. a) **Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?**

A melhor opção para menores riscos seria o bairro Brooklyn. Além de ter um preço médio de $123.50, é o bairro com a menor disponibilidade no ano, o que indica uma alta procura.

```
Localização: Manhattan -> {'mean_value': '194.68', 'mean_review': '20.99', 'mean_dispo': '111.98'}
Localização: Brooklyn -> {'mean_value': '123.50', 'mean_review': '24.20', 'mean_dispo': '100.22'}
Localização: Queens -> {'mean_value': '97.77', 'mean_review': '27.70', 'mean_dispo': '144.45'}
Localização: Staten Island -> {'mean_value': '114.81', 'mean_review': '30.94', 'mean_dispo': '199.68'}
Localização: Bronx -> {'mean_value': '87.50', 'mean_review': '26.00', 'mean_dispo': '165.76'}
```

2. b) **O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?**

Não, o número de reviews e a disponibilidade podem afetar o fator "popularidade" do imóvel, mas não é seguro dizer que influenciam o preço, uma vez que tanto os dados com valores altos quanto os com valores baixos apresentam diversas variações nessas duas variáveis.

```
----------Group: Manhattan Room Type: Entire home/apt----------

Max Price:
                                    nome bairro_group          bairro       room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
29661  East 72nd Townhouse by (Hidden by Airbnb)    Manhattan  Upper East Side  Entire home/apt              1                  0  7703.0                  146

Min Price:
                                        nome bairro_group          bairro       room_type  minimo_noites  numero_de_reviews  price  disponibilidade_365
2859   Large furnished 2 bedrooms- - 30 days Minimum    Manhattan     East Village  Entire home/apt             30                  0   10.0                  137
23255        Quiet, Cozy UES Studio Near the Subway    Manhattan  Upper East Side  Entire home/apt              3                 10   10.0                    0

Mean Price:  246.40
```

```
----------Group: Manhattan Room Type: Private room----------

Max Price:
                                nome bairro_group          bairro     room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
37193  Apartment New York \nHell's Kitchens    Manhattan  Upper West Side  Private room             30                  0  6500.0                   97

Min Price:
                                       nome bairro_group          bairro     room_type  minimo_noites  numero_de_reviews  price  disponibilidade_365
22286                                  Jen Apt    Manhattan            SoHo  Private room              5                  2   10.0                    0
31065  Very Spacious bedroom, steps from CENTRAL PARK.    Manhattan  Upper West Side  Private room              1                  2   10.0                    0
31406                     Cozy feel at home studio    Manhattan        Kips Bay  Private room              5                 42   10.0                    2

Mean Price:  115.54
```

2. c) **Existe algum padrão no texto do nome do local para lugares de mais alto valor?**

Sim, como comentado no início, existe uma clara relação entre o texto do anúncio e o valor a ser pago. Descrições que mencionam mansões, apartamentos luxuosos, barcos, locais de gravação de filmes, casas e nomes de eventos tendem a estar associadas a anúncios de maior valor.

```
----------Group: Bronx Room Type: Shared room----------

Max Price:
                                    nome bairro_group     bairro    room_type  minimo_noites  numero_de_reviews  price  disponibilidade_365
25421  New York's Hidden Secret for luxury living       Bronx  Riverdale  Shared room              2                  1  800.0                  269
```

```
----------Group: Bronx Room Type: Private room----------

Max Price:
                        nome bairro_group     bairro     room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
24476  "The luxury of Comfort"       Bronx  Riverdale  Private room              2                  0  2500.0                  179
```

```
----------Group: Staten Island Room Type: Entire home/apt--------------------

Max Price:
                         nome   bairro_group        bairro       room_type  minimo_noites  numero_de_reviews   price  disponibilidade_365
22352  Victorian Film location  Staten Island  Randall Manor  Entire home/apt              1                  0  5000.0                  344
```

3. **Explique como você faria a previsão do preço a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?**

Devido a irregularidade dos dados, como apresentado anteriormente, optei por utilizar o modelo de regressão não linear que apresenta vantagens em modelos sem padrão para os dados, a vantagem do modelo não linear é que não assume relações lineares entre variáveis e tem uma alta capacidade ao lidar com categorias porém, uma grande desvantagem desse modelo é sua lentidão quanto ao processamento.

O problema a ser resolvido é a regressão, pois é utilizada para prever valores específicos, neste caso, o preço. Comquanto, considerando a maximazação do aprendizado do modelo optei por utilizar **ano, o preço, latitude, longitude e tipo de acomodação** para ter uma previsão mais segura.

Código usado:

```
def future_price_prediction():
    data_correct = table[table['year'].notna()]

    data_meam = round(data_correct['year'].mean())
    prices_correct = table[table['price'] < 8000]

    corrected_meam = round(prices_correct['price'].mean())
    table_corrected = table.copy()
    table_corrected.loc[table_corrected['price'] >= 8000, 'price'] = corrected_meam
    table_corrected['year'] = table_corrected['year'].fillna(data_meam)
    features = ['bairro_group', 'bairro', 'room_type', 'year']

    target = 'price'
    table_corrected = table_corrected.dropna(subset=features + [target])
    x = table_corrected[features]

    y = table_corrected[target]
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
    cat_cols = ['bairro_group', 'bairro', 'room_type']

    num_cols = ['year']
```

```
preprocessor = ColumnTransformer(transformers=[('cat', OneHotEncoder(handle_unknown='ignore'), cat_cols),('num', 'passthrough', num_cols)])

pipeline = Pipeline(steps=[('preprocessing', preprocessor),('model', RandomForestRegressor(n_estimators=100, random_state=42, n_jobs=1))])
  pipeline.fit(x_train, y_train)
  predicted_prices = pipeline.predict(x)
  table_corrected['predicted_price'] = np.round(predicted_prices, 2)
  return table_corrected[['nome', 'bairro_group', 'price', 'year', 'predicted_price']]
print(future_price_prediction())
```

	nome	bairro_group	price	year	predicted_price
0	Skylit Midtown Castle	Manhattan	225.0	2019.0	310.49
1	THE VILLAGE OF HARLEM....NEW YORK !	Manhattan	150.0	2018.0	88.45
2	Cozy Entire Floor of Brownstone	Brooklyn	89.0	2019.0	200.33
3	Entire Apt: Spacious Studio/Loft by central park	Manhattan	80.0	2018.0	166.46
4	Large Cozy 1 BR Apartment In Midtown East	Manhattan	200.0	2019.0	266.67
...	...	...	...	...	...
48889	Charming one bedroom - newly renovated rowhouse	Brooklyn	70.0	2018.0	84.11
48890	Affordable room in Bushwick/East Williamsburg	Brooklyn	40.0	2018.0	62.39
48891	Sunny Studio at Historical Neighborhood	Manhattan	115.0	2018.0	192.25
48892	43rd St. Time Square-cozy single bed	Manhattan	55.0	2018.0	166.84
48893	Trendy duplex in the very heart of Hell's Kitchen	Manhattan	90.0	2018.0	148.76

4. **Supondo um apartamento com as seguintes características: Qual seria a sua sugestão de preço?**

Considerando o modelo escolhido, o valor sugerido é de $318.77. Para este resultado adaptei o código de previsão para que filtrasse pela latitude e longitude esperados.

```
Previsão do preço para a localização (40.7488, -73.9855) em 2025: $318.77
```