Entregas

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

   No Jupyter Notebook foi realizada uma análise de outra base de dados retirada do ourhome.nyc/ com diversos indicadores sociais por bairros e distritos da cidade de Nova Iorque. Entretanto, os distritos contidos no senso e no dataset não batiam, me faltou conhecimento técnico para utilizar a latitude e longitude dos apartamentos para descobrir o zipcode e adequar ao distrito e seus indicadores. Quando testei agregar os datasets, apesar de um score muito baixo nos modelos, foi notado que fatores como desemprego, desigualdade e renda média são muito relevantes. Infelizmente, agregar os indicadores pelas médias nos grandes bairros generalizou demais as informações.
2. Responda também às seguintes perguntas:
   a. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra? Manhattan
   b. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço? Menos do que outros fatores, como calculado_host_listing_count, bairro e tipo de apartamento.
   c. Existe algum padrão no texto do nome do local para lugares de mais alto valor? Percebi principalmente adjetivos que remetem ao luxo e privacidade *(luxury, cozy, quiet, private)*
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?

Tentei utilizar o máximo de colunas possíveis, transformando algumas classificadas categoricamente por numericamente, como o tipo de quarto, substituindo os tipos por ordem decrescente de valor agregado a apartamentos inteiros, quartos privativos e quartos compartilhados. Também adaptei o modo em que as reviews estavam cadastradas, para uma melhor interpretabilidade da máquina. Utilizei modelos de aprendizado de máquina de regressão tanto por regressão linear, como por uma Random Forest. Os prós são que após o tratamento dos dados realizado de uma maneira minuciosa ele normalmente traz bons resultados. Como contras, acredito que não foi a melhor escolha para analisar apartamentos, ou então dados de apartamentos cadastrados pelos seus locadores, que nem sempre contam com detalhes realmente relevantes para o preço, como se ele está mobiliado, quantidade de banheiros, quartos e entre outros parâmetros que poderiam ter melhorado a performance do modelo.

4. Supondo um apartamento com as seguintes características:

```
{'id': 2595,
 'nome': 'Skylit Midtown Castle',
 'host_id': 2845,
 'host_name': 'Jennifer',
 'bairro_group': 'Manhattan',
 'bairro': 'Midtown',
 'latitude': 40.75362,
 'longitude': -73.98377,
 'room_type': 'Entire home/apt',
 'price': 225,
 'minimo_noites': 1,
 'numero_de_reviews': 45,
 'ultima_review': '2019-05-21',
 'reviews_por_mes': 0.38,
 'calculado_host_listings_count': 2,
 'disponibilidade_365': 355}
```

Qual seria a sua sugestão de preço?

$226,28