Case Técnico Programa Lighthouse – Aurelio Guilherme

Link: Streamlit

Para responder as perguntas utilizei a aplicação do Streamlit, que contem presentes as seguintes análises na etapa de Data Understanding:

- Análise Descritiva
  - Dados Faltantes
  - Dados Duplicados
  - Análise de Dados Contínuos
  - Valores Discrepantes
  - Exploração Individual das Variáveis Contínuas
    - price
    - minimo_noites
    - numero_de_reviews
    - reviews_por_mes
    - calculado_host_listings_count
    - disponibilidade_365
  - Análise de Dados Categóricos
    - host_id
    - bairro_group
    - bairro
    - room_type
  - Exploração Multivariada
    - Correlação
  - Analisando a relação entre variáveis
    - Relação de preço com os bairros
    - Relação de preço com o tipo de espaço

        - Relação de preço com o mínimo de noites de forma agrupada
        - Relação de preço com o numero de reviews de forma agrupada
        - Relação de preço com o a quantidade de imoveis por host de forma agrupada
        - Relação de preço com o a variável disponibilidade_365 de forma agrupada
    - Dados Geoespaciais
    - Análise Temporal
    - Análise de Dados Textuais.
        - Comparação de Palavras-Chave por Faixa de Preço
        - Palavras presentes somente em imóveis de alto valor

Análise Inferencial

    - Hipóteses
        - Existe diferença entre os preços dos imóveis de acordo com o tipo de espaço?
        - Existem diferenças significativas entre o preço nos grupos de bairros?
        - Existe associação entre o tipo de espaço e o bairro?

Perguntas de negócio:

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

Resposta: Considerando o grupo de bairros, o mais popular e rentável é Manhattan, deste modo é possível efetuar as seguintes análises em relação aos bairros presentes nesse grupo:

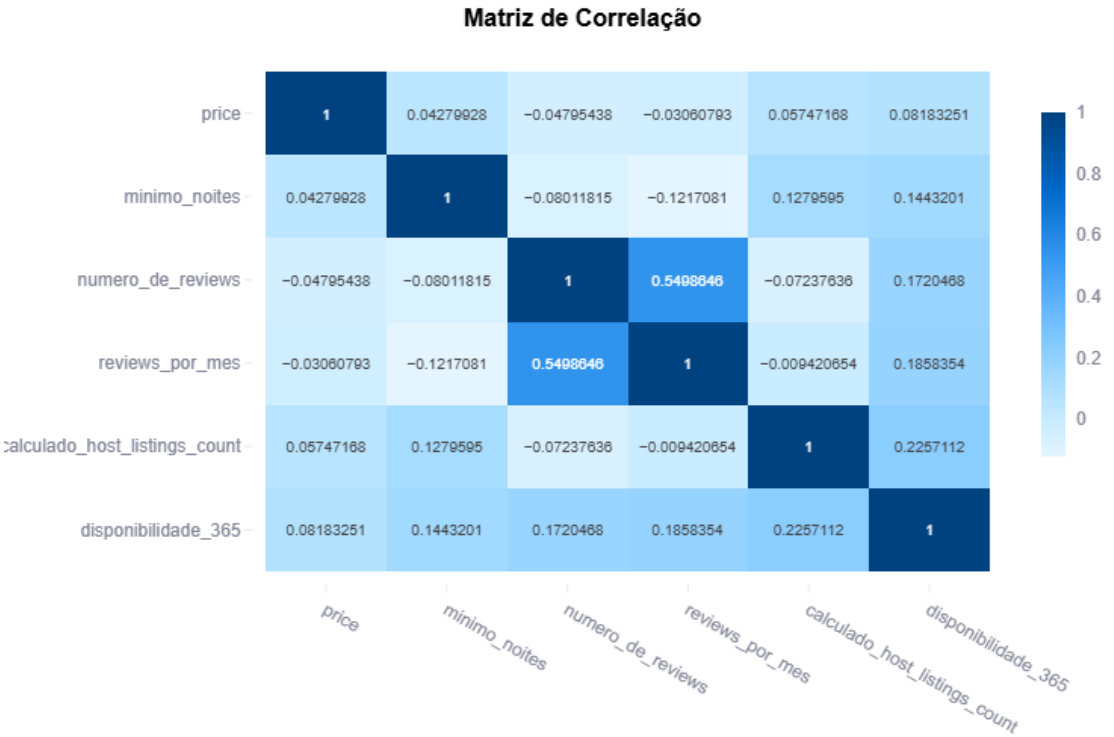
- Midtown possui o maior valor acumulado de diárias, US$436.801,00 e também corresponde a 6º posição referindo-se à popularidade anúncios e a média de preço em seu grupo de bairros, com 1545 anúncios e US$ 282,71 respectivamente.

Embora existam bairros mais populares presentes no grupo do Brooklyn como, por exemplo, Williamsburg, que inclusive possui um valor acumulado superior a Midtown, esses outros bairros não tem uma média tão elevada.

Portanto, eu indicaria a compra do imóvel no bairro Midtown por estar entre os mais populares e com uma das maiores médias.

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

Resposta: As variáveis não apresentam correlação significativa em preço.



Agrupando as variáveis em grupos, é possível analisar que os grupos se comportam de formas distintas, tendo médias diferentes e soma distintas.

minimo_noites_grupo	count	% do total	mean	sum	min	max
Entre 1 a 3 dias	32407	66.30	148.085537	4799008	10	8000
Entre 1 e 2 Semanas	1433	2.93	126.753664	181638	15	6000
Entre 2 Semanas e 1 Mes	5151	10.54	170.768006	879626	10	10000
Entre 3 a 7 dias	9145	18.71	156.197485	1428426	10	10000
Mais de 1 Ano	14	0.03	137.571429	1926	45	400
Mais de 1 Mes	415	0.85	188.918072	78401	24	4200
Mais de 2 Meses	251	0.51	331.756972	83271	25	10000
Mais de 6 Meses	67	0.14	221.388060	14833	12	2350

Figura 1Tabela de agrupamento por mínimo de noites

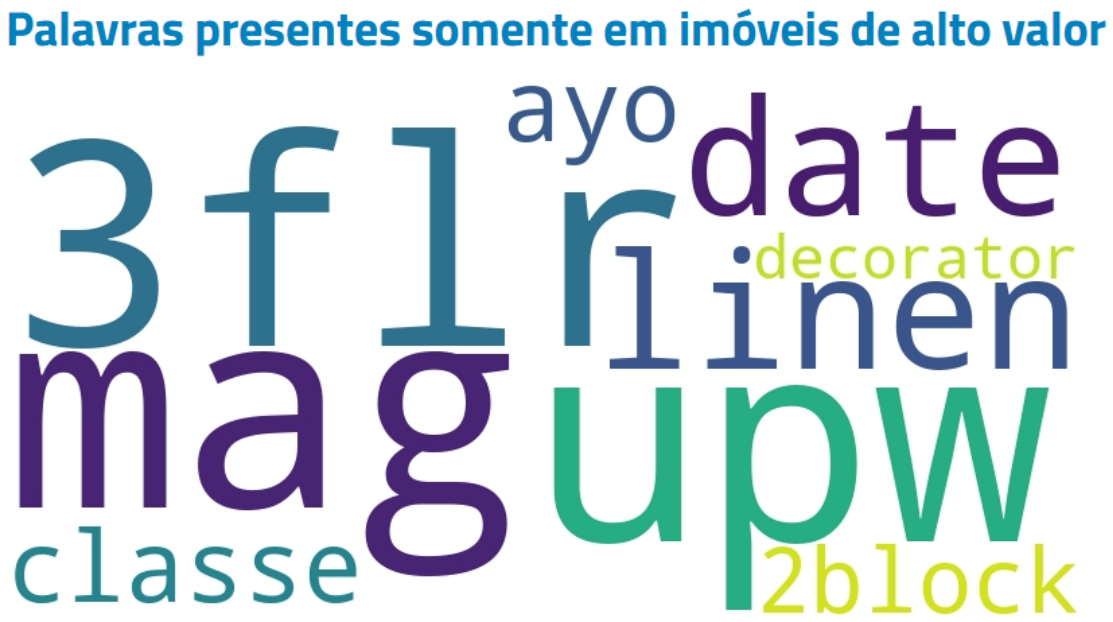
disponibilidade_365_grou	count	% do total	mean	sum	min	max
0 Dias	17530	35.86	136.055391	2385051	10	10000
1 Ano	1294	2.65	250.848532	324598	20	9999
Entre 1 a 3 dias	984	2.01	136.189024	134010	10	1150
Entre 1 e 2 Semanas	1271	2.60	140.296617	178317	13	1680
Entre 2 Semanas e 1 Mes	1981	4.05	142.294801	281886	20	3000
Entre 3 a 7 dias	1037	2.12	132.237223	137130	10	840
Mais de 1 Mes	2996	6.13	137.826435	412928	10	6419
Mais de 2 Meses	8724	17.85	157.535763	1374342	10	10000
Mais de 6 Meses	13066	26.73	171.350605	2238867	10	8500

Figura 2 Tabela de agrupamento por disponibilidade 365

3. Existe algum padrão no texto do nome do local para lugares de mais alto valor?

Respota: Analisando a nuvem de palavras da divisão entre imóveis de baixo padrão com os imóveis de alto padrão divididos pela mediana, é possível identificar ao contar a frequência entre os dois grupos, ranqueá-las e subtrair, mantendo assim somente as palavras que estão presentes em um dos grupos.

No caso, os imóveis de alto padrão possuem indicação de que ficam localizados no distrito de Manhattan, ou próximas a algum ponto de interesse.



4 - 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?

Resposta: Utilizei uma abordagem de experimentação, tanto para features, quanto para modelos em um problema de regressão, foram utilizados 3 modelos diferentes e o que teve melhor performance foi o LinearRegression.

Para possibilitar a experimentação, utilizei o MLFlow e criei uma classe que possibilita a adição e remoção de features de forma simples, testei dois conjuntos de features, um dos conjuntos efetuei um tratamento para transformar as algumas variáveis de forma categórica e o outro mantive as features numéricas padronizadas, assim pude aferir a performance desse teste, porem ambos os modelos demonstraram-se ineficazes, o melhor modelo obteve o resultado de R^2 0.10 que indica o modelo não consegue explicar a variação dos dados, e também um RMSE 214,55 ou seja, o modelo não faz boas predições, errando em média US$214,55

O registro desses experimentos ficou registrados e podem ser conferidos no MLFlow.

A relação utilizada nos experimentos são:

Experimento 1:

Features_numericas

- numero_de_reviews
- reviews_por_mes
- calculado_host_listings_count

features_categoricas

- room_type
- bairro_group
- minimo_noites_categorico
- disponibilidade_365_categorico
- ultima_review_semestre

- valor_preenchido

Experimento 2:

features_numericas

- numero_de_reviews
- reviews_por_mes
- calculado_host_listings_count
- disponibilidade_365
- latitude
- longitude
- minimo_noites

features_categoricas

- room_type
- bairro_group
- valor_preenchido

Agora em uma abordagem futura, pensando exclusivamente em garantir um modelo mais performático, efetuaria um tratamento adicional nos valores outliers, deste modo um modelo simples, como Regressão Linear, possivelmente teria melhor performance do que os até então experienciados.

5. 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',

'minimo_noites': 1,

'numero_de_reviews': 45,

'ultima_review': '2019-05-21',

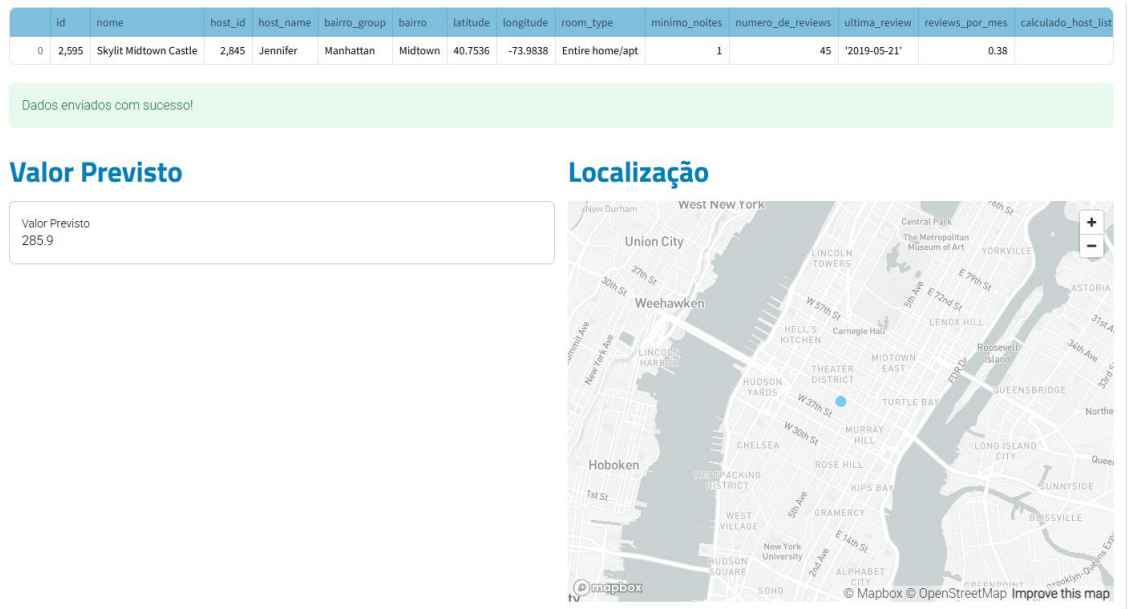
'reviews_por_mes': 0.38,

'calculado_host_listings_count': 2,

'disponibilidade_365': 355}

Qual seria a sua sugestão de preço?

O valor previsto foi de US$ 285.90



6 – Link do Vídeo.