

Seazone
Processo Seletivo
Desafio Data Analysis

Análise de Dados Imobiliários

Relatório referente ao desafio dos anúncios imobiliários
do Airbnb.

Autor: Fabricio Cravo

29/05/2021

Conteúdo

1	Descrição dos Dados	1
2	Metodologia para análise	3
3	Tarefa 1	3
4	Tarefa 2	5
5	Tarefa 3	8
6	Tarefa 4	10
7	Conclusão e Feedback	11

1 Descrição dos Dados

No desafio foram entregues duas tabelas uma chamada *desafio details* e a outra *desafio priceav*. A tabela *desafio details* contém as informações referentes ao anúncio de um imóvel, como um identificador, o bairro, a nota do anúncio, a indicação de superhost, a quantidade de quartos e banheiros e o título. Uma visão inicial do dataset pode ser vista na figura 1.

[36]:

Unnamed: 0	airbnb_listing_id	suburb	ad_name	number_of_bedrooms	number_of_bathrooms	star_rating	is_superhost	number_of_reviews	
0	0	31389869	Jurerê	Lindo Apartamento em Jurerê	2.0	2.0	5.0	False	15.0
1	1	40010667	Canasvieiras	Residencial Arruda, 1 quarto	1.0	1.0	NaN	False	0.0
2	2	38905997	Ingleles	Apartamento NOVO Completo - Moderno e Sofisticado	1.0	1.0	4.5	True	13.0
3	3	22343656	Ingleles	06- Apartamento 02 habitaciones	2.0	1.0	5.0	True	28.0
4	4	18328184	Canasvieiras	Apto 2 quartos em Canasvieiras, Florianopolis!	2.0	1.0	5.0	True	35.0

Figura 1: Visualização inicial da tabela *desafio details*

Devemos observar a quantidade de dados faltando de cada atributo na tabela. Essa observação revela a qualidade dos dados, já que dados incompletos podem afetar os cálculos e os resultados. No total temos 4691 linhas na tabela e podemos ver a quantidade de elementos faltantes na figura 2. Nessa figura *True* representa um dado faltante. Assim podemos perceber que a maioria dos dados está quase completa, com a exceção do número de estrelas por anúncio (*star review*) onde temos 2121 linhas faltantes na tabela. Será necessário tratar esses dados para conseguir valores confiáveis dos cálculos.

```
Unnamed: 0 {False: 4691}
airbnb_listing_id {False: 4691}
suburb {False: 4691}
ad_name {False: 4691}
number_of_bedrooms {False: 4508, True: 183}
number_of_bathrooms {False: 4690, True: 1}
star_rating {False: 2570, True: 2121}
is_superhost {False: 4691}
number_of_reviews {False: 4684, True: 7}
```

Figura 2: Dados faltantes da tabela *desafio details*, *True* representa a falta

A tabela *desafio priceav* contém as informações referentes aos valores dos anúncios de um imóvel e o faturamento. Infelizmente, nessa tabela não existem um id para cada reservaçao diferente. Como não foi especificado no desafio os dados foram tratados da seguinte maneira, foi considerado que o preço é diário e que o faturamento existe somente quando o anúncio é

listado como ocupado. Considera-se também que uma reserva pode ocupar diversas linhas na tabela então considere o dia que a reserva foi feita e o id do anúncio para distinguir entre reservas iguais presentes em mais de uma linha da tabela. A visualização inicial dessa tabela se encontra na figura 3. Podemos também ver que não existem dados faltantes na figura 4.

[37]:

	Unnamed: 0	Unnamed: 0.1	airbnb_listing_id	booked_on	date	price_string	occupied
0	0	2148	40201349	blank	2020-11-15	250.0	0
1	1	2159	40201349	blank	2020-11-26	250.0	0
2	2	2160	40201349	blank	2020-11-27	250.0	0
3	3	2173	40201349	blank	2020-12-10	250.0	0
4	4	2226	40201349	blank	2021-02-01	250.0	0

Figura 3: Visualização inicial da tabela *desafio priceav*

```

Unnamed: 0 {False: 4691}
airbnb_listing_id {False: 4691}
suburb {False: 4691}
ad_name {False: 4691}
number_of_bedrooms {False: 4508, True: 183}
number_of_bathrooms {False: 4690, True: 1}
star_rating {False: 2570, True: 2121}
is_superhost {False: 4691}
number_of_reviews {False: 4684, True: 7}

```

Figura 4: Dados faltantes da tabela *desafio priceav* , *True* representa a falta

2 Metodologia para análise

Para uma boa análise dos dados é essencial a comunicação com o cliente e o entendimento do seu objetivo final ou do problema a ser resolvido. Só assim um analista pode julgar se as métricas e os dados são suficientes para resolver o problema. Como o objetivo não foi especificado no desafio a análise será generalista. Os dados foram processados em Python usando a biblioteca Pandas e um Jupyter Notebook foi feito para descrever em detalhes o processamento dos dados.

3 Tarefa 1

Nessa tarefa foi pedido para listar o número de anúncios por bairro. Essa métrica é interessante para visualizar a oferta de cada região. Na figura 5 podemos visualizar o número bruto de anúncios por bairro e na figura 6 podemos visualizar o gráfico de barras que eles produzem. Desses dados podemos perceber que o bairro dos Ingleses é o bairro com maior oferta é o dos Ingleses com a menor oferta presente no Centro. Os dados da oferta podem ser relacionados com os dados da demanda para tomar conclusões estratégicas referentes a construção de novos imóveis na região.

```
Ingleses 2388
Canasvieiras 1177
Jurerê 539
Lagoa da Conceição 309
Centro 278
```

Figura 5: Listings por bairro

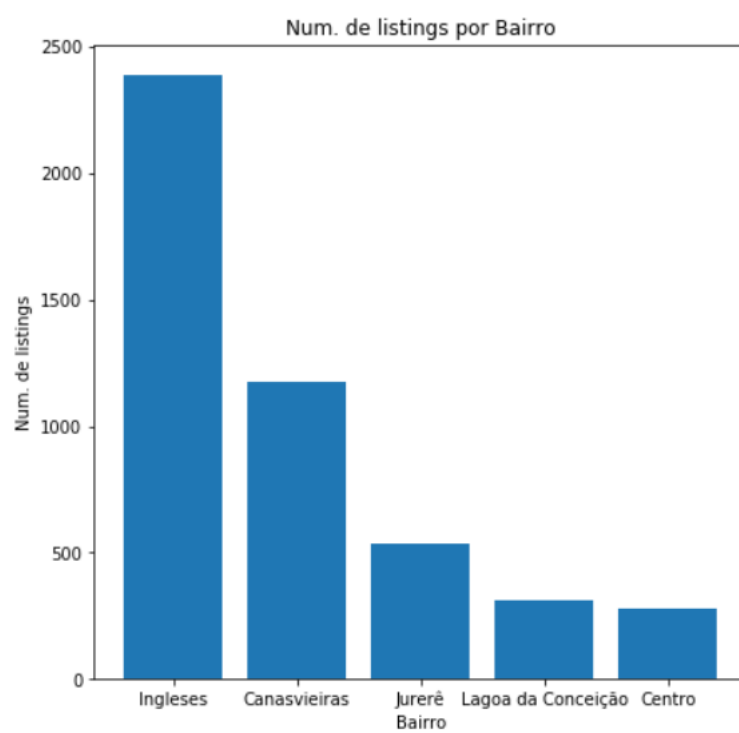


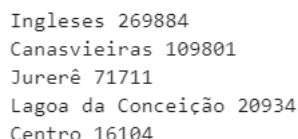
Figura 6: Gráfico de barras listings por bairro

4 Tarefa 2

Para realizar essa tarefa começou-se calculando o faturamento médio por listing e adicionando a tabela *desafio details*. Para o cálculo do faturamento médio foram somados os preços de todos os dias que o imóvel estava ocupado e dividido pelos dias que estava disponível. Isso foi feito diminuir o faturamento médio de imóveis que são raramente alugados e que ficam muitos dias disponíveis, ou seja, não recebem ofertas. Ainda, foram removidos do cálculo os imóveis que não estão listados em *desafio priceav* mas estão em *desafio details*. Temos ainda que o número de ids na tabela *desafio details* que não está listado em *desafio priceav* é de 1344 dados, em torno de 25% do número de ids. Então essa ação claramente tem um impacto profundo no resultado.

Na figura 7 temos a soma do faturamento médio para todos os imóveis listados por bairro e na figura 8 temos o gráfico de barras respectivo a esses dados.

Infelizmente a soma de todos os faturamentos médios pode não ser um bom representativo do lucro da região. Certas regiões possuem um número maior de imóveis e podem acabar ganhando nessa medida devido a seu grande número de imóveis. Assim decidi normalizar o valor do faturamento bruto dividido pelo número de listagens presentes para cada bairro. Os resultados normalizados podem ser encontrados na figura 9 e na figura 10.



Ingleses	269884
Canasvieiras	109801
Jurerê	71711
Lagoa da Conceição	20934
Centro	16104

Figura 7: Soma de todos os faturamentos médios por bairro

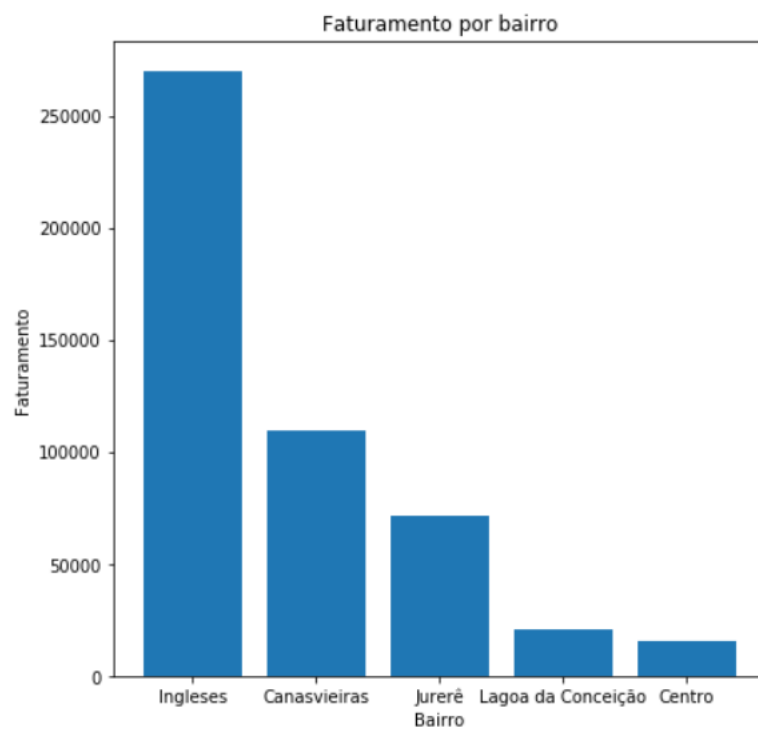


Figura 8: Gráfico de barras da soma dos faturamentos médios por bairro

Jurerê 192
 Ingleses 157
 Canasvieiras 129
 Lagoa da Conceição 96
 Centro 81

Figura 9: Faturamento normalizado pelo número de listings

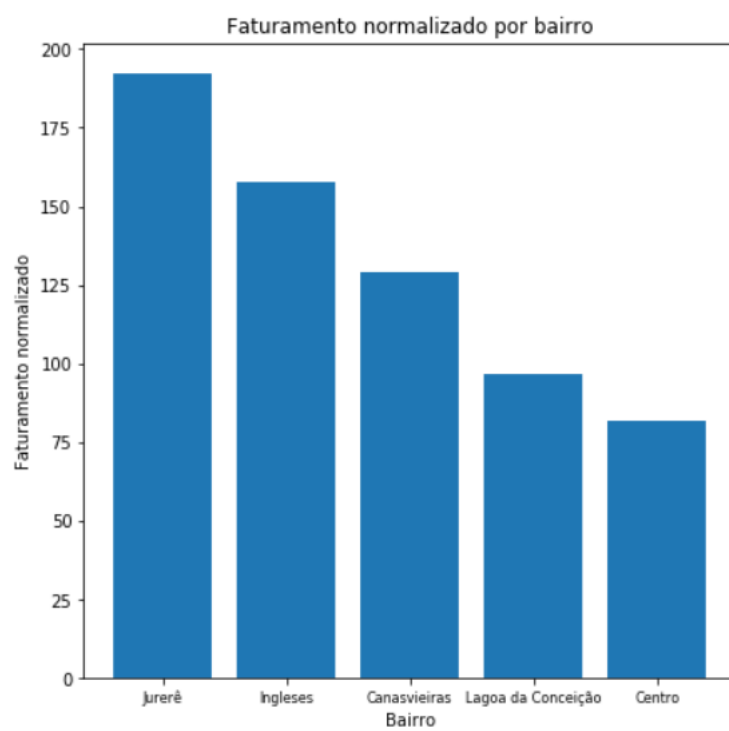


Figura 10: Gráfico de barras do faturamento normalizado pelo número de listings

5 Tarefa 3

Nessa tarefa é necessário calcular a correlação das características do anúncio com seu faturamento médio. Para obter bons resultados substituímos os bairros da tabela por valores numéricos ordem crescente de 1 a 5 igual a obtida na figura 10 com os bairros vencedores recebendo um valor maior. Isso é feito para ajudar os cálculos de correlação já que espera-se que os bairros com maior faturamento médio normalizado tenham faturamentos médios mais altos. Os resultados do cálculo da correlação podem ser encontrados na figura 11 e na figura 12.

Podemos perceber nessa figura que as características com a correlação mais alta são o número de banheiros, quartos e o bairro. Isso é esperado, pois bairros mais nobres e casas maiores geralmente cobram mais caro por seu aluguel, gerando um faturamento médio bruto maior se existir uma demanda para elas. Também era esperado que o número de reviews e o fato de o host ser qualificado como super host ou não, teriam uma correlação pequena com o faturamento. Isso se deve ao fato que esses fatores não são tão determinantes quanto o preço e as necessidades do consumidor na escolha do imóvel. Contudo era esperado que o número de estrelas possuísse um impacto maior no faturamento, existem diversos motivos que podem ter causado esse dado não esperado. O primeiro pode ser devido ao fato de ser a característica com dados mais ausentes de todas. Eu recomendaria uma análise com mais dados e um cálculo da média do faturamento de cada quartil do número de estrelas para conferir se o resultado é realmente confiável.

```
airbnb_listing_id    -0.000042
suburb               0.218078
number_of_bedrooms   0.292396
number_of_bathrooms  0.272471
star_rating          0.025349
is_superhost         -0.047013
number_of_reviews    -0.035029
Name: Faturamento, dtype: float64
```

Figura 11: Correlação entre os diferentes aspectos do anúncio com o faturamento médio

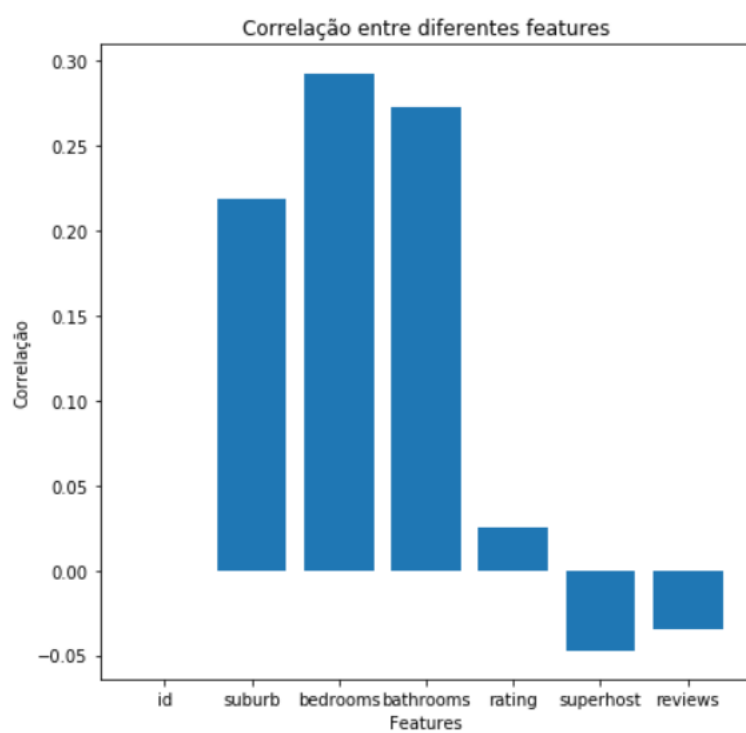


Figura 12: Gráfico de correlação entre os diferentes aspectos do anúncio

6 Tarefa 4

Para essa tarefa é necessário calcular a antecedência média da reserva e os dias que os imóveis são ocupados. O dataset *desafio priceav* não possui uma variável referente a um identificador da reserva, então usamos como identificador da reserva o dia que o imóvel foi alugado e o id do listing (mais detalhes do cálculo no Jupyter Notebook). Com esse cálculo foi obtida uma média de 20 dias de antecedência na data de reservação antes da ocupação do imóvel. Também foi analisado a ocupação por dia da semana e os resultados podem ser encontrados na figura 13. Podemos perceber por esse gráfico que os dias da semana vencedores são a sexta, sábado e domingo.

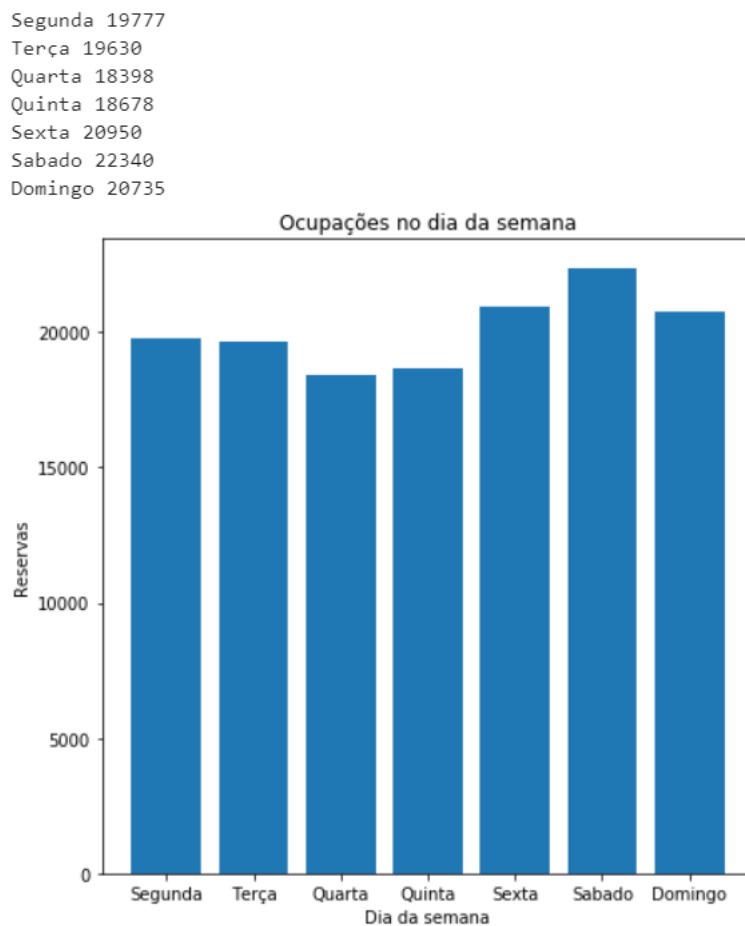


Figura 13: Ocupação por dia da semana

7 Conclusão e Feedback

Primeiramente, achei o desafio muito interessante. Gosto muito da análise de dados e achei extremamente pertinente a aplicação no mercado imobiliário. Peço desculpas pela falta de experiência no setor, acredito que isso poderia enriquecer muito a qualidade das minhas conclusões sobre os dados.

Espero que a análise atinja o nível esperados pela empresa. Infelizmente, sem saber o objetivo desejado é difícil fazer uma observação aprofundada e responder se a informação extraída pelos dados gera o conhecimento que empresa deseja ter para tomar decisões. Seria também de muita ajuda ter acesso a maneira que os dados foram gerados e coletados, pois isso pode explicar os resultados obtidos e permitir conclusões sobre viés de amostragem.

Muito obrigado pela leitura e pela oportunidade.