Deliverable 1

Roof Imóveis

Aliny Pautz Sunderhus

Projeto submetido em cumprimento parcial
dos requisitos para a conclusão do curso
Data Expert da Escola DNC.

São José dos Campos - SP - Brazil
Agosto, 2022

Onde encontrar o código

O código desenvolvido para esse projeto encontra-se em um notebook Colab hospedado neste repositório público Github.

1 Questão do negócio

Este projeto se propõe a realizar uma análise exploratória de um conjunto de dados referente a vendas de imóveis no Condado de King.

Sendo o condado mais populoso dos 39 existentes no estado de Washington (EUA), o Condado de King tem na imigração a causa de mais da metade do crescimento populacional experimentado nas últimas décadas. Diferentemente de outros condados em que a maioria da população está concentrada na capital, a maior parte dos habitantes do Condado de King mora em cidades vizinhas e outras áreas urbanas. Menos de um terço reside na capital Seattle, apesar desta ser a cidade mais populosa de King. Maiores informações acerca da região podem ser encontradas na página oficial do Condado de King.

Para a resolução deste projeto, considera-se a empresa fictícia Roof Imóveis. Esta empresa do ramo imobiliário deseja expandir sua área de atuação e investir internacionalmente.

O principal objetivo desta consultoria é selecionar 5 imóveis do Condado de King em que a Roof Imóveis deveria investir e 5 dos quais o investimento não é recomendado.

Para alcançar essa meta, pretende-se criar métricas para a classificação dos imóveis cujos dados estão disponíveis.

2 Entendimento do negócio

O Dataset a ser analisado está disponível nesta página do Kaggle, uma comunidade online de cientista de dados.

O conjunto de dados contém 21613 registros de vendas de imóveis, realizadas entre maio de 2014 e maio de 2015 no Condado de King. Estes imóveis são do tipo casas residenciais.

Os registros possuem 21 características, sendo elas:

- id: identificador do imóvel;
- date: data de venda;
- price: preço de venda;
- bathroomss: número de banheiros;
- bedrooms: número de quartos;
- sqft _living: área habitável, medida em $pés^2$;
- sqft_lot: área do lote, medida em $pés^2$;
- floors: número de andares;
- waterfront: recebe o valor '1' se a propriedade é beira-mar e '0' se não for;
- view: índices de 0 a 4 indicando a quão boa a vista do imóvel é;
- condition: condição do imóvel, classificada de 0 a 5;
- grade: classificação por qualidade de construção que se refere aos tipos de materiais utilizados e à qualidade de acabamento;
- sqft_above: área acima do solo, medida em $pés^2$;
- sqft_basmt: área abaixo do solo, medida em $pés^2$;
- yr_built: ano da construção;
- yr_renov: ano da reforma - recebe '0' caso o imóvel nunca tenha sido reformado;
- zipcode: código postal, equivalente ao CEP brasileiro;
- lat: latitude;
- long: longitude;
- sqft_liv15: área habitável média dos 5 imóveis mais próximos, medida em $pés^2$;

- sqft_lot15: área média dos lotes dos 5 imóveis mais próximos, medida em $pés^2$.

Para análise dos imóveis, os atributos mais utilizados foram:

- área habitável do imóvel;
- qualidade da construção;
- área habitável média dos 15 imóveis mais próximos;
- cidade em que o imóvel está localizado, obtida a parir do zip code.

3 Coleta de dados

O arquivo que continha os dados era do tipo CSV (*Comma-separated values*). Este foi adicionado a um repositório na plataforma Github e, utilizando a biblioteca Pandas, os dados puderam ser carregados utilizando o link em que o arquivo está disponível, como no código abaixo:

```
url = 'url_dos_dados'
df = pd.read_csv(url)
```

Essa estratégia foi adotada para evitar que o *upload* do arquivo CSV tivesse de ser efetuado toda vez que o notebook Google Colab fosse reiniciado.

A partir dessa análise inicial dos dados possível concluir que:

- O conjunto de dados não possui valores nulos;
- os dados da coluna *date* deveriam ser do tipo *datetime*;
- os dados da coluna *bathrooms* deveriam ser do tipo inteiro;
- não há linhas duplicadas;
- há 177 registros de identificadores *id* duplicados. Isso indica que um mesmo imóvel foi vendido mais de uma vez. Através dessas informações, podemos averiguar se houve lucro ou prejuízo na segunda venda do imóvel;

- a quantidade máxima de quartos é 33, um valor muito acima da média de todos os registros. Essa questão será verificada.
- a maioria (99,25%) dos registros não são do tipo beira-mar;
- a maioria dos registros (90,17%) possuem zero como valor para a vista;
- cerca de 6% dos registros não possuem porão;
- para melhor entendimento do cliente, seria melhor que os nomes dos atributos fossem em português;
- ao pensarmos sobre localização dos imóveis, é interessante sabermos em quais cidade esses se encontram;
- ao avaliar o preço de um imóvel, é interessante utilizarmos uma métrica de preço/área.

4 Limpeza dos dados

Nesta seção, os dados de data da construção do imóvel foram convertidos para o tipo *datetime* e os dados de número de banheiros foram arredondados e convertidos para o tipo inteiro.

Foi removido o registro em que o número de quartos era 33, por ter sido considerado um *outlier*. As colunas foram renomeadas para português, que é a língua do cliente.

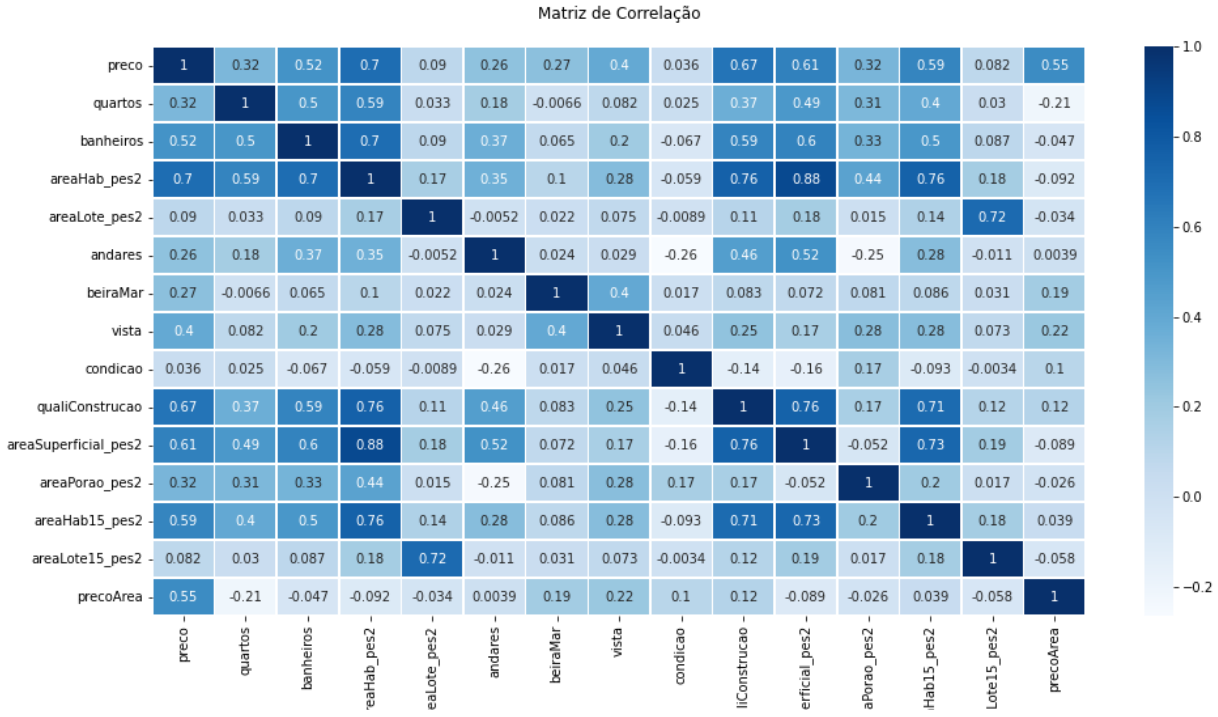
Além disso, foi criada uma função que busca a cidade dos Estados Unidos referente ao *zip code* do imóvel e essa informação foi adicionada a uma nova coluna chamada "cidade".

Por fim, uma nova coluna foi criada para receber a métrica *preço/área habitável*.

5 Exploração dos dados

Primeiramente foi criada uma matriz de correlação para averiguar quais os atributos numéricos presentes no conjunto de dados estão mais

correlacionados com o preço dos imóveis. A matriz obtida encontra-se na figura a seguir.



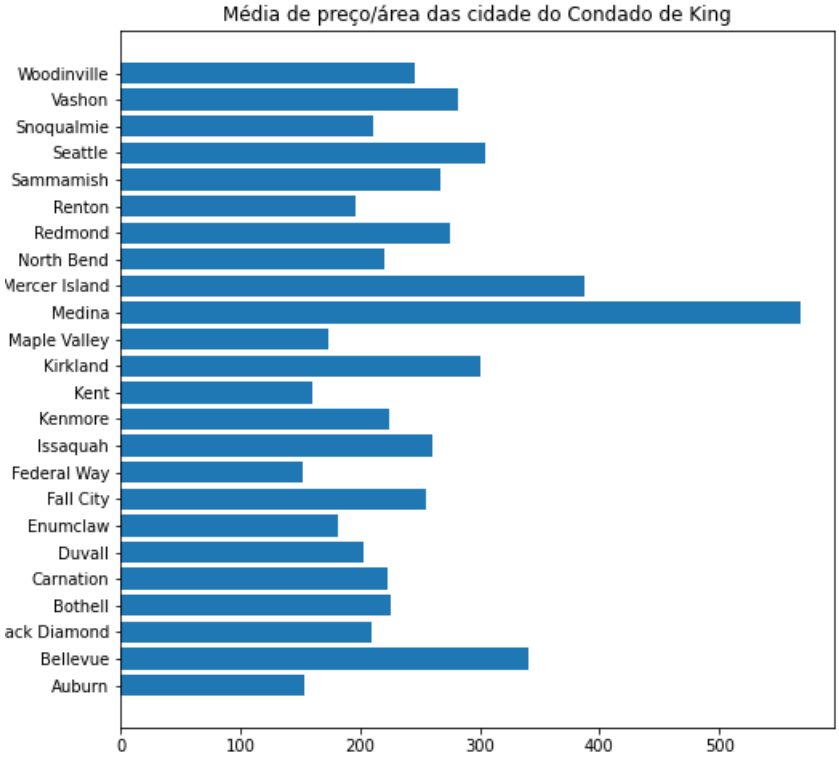
As características que se mostraram mais correlacionadas com o preço do imóvel foram:

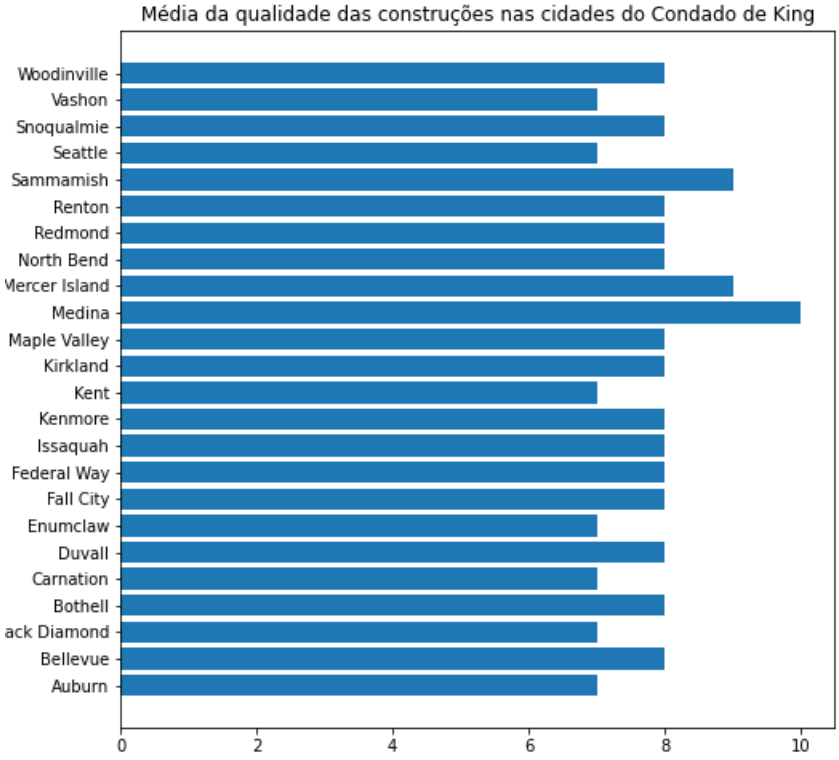
- Área habitável do imóvel (*areaHab_pes2*);
- qualidade dos materiais utilizados na construção (*qualiConstrucao*);
- área habitável média dos 15 imóveis mais próximos (*areaHab15_pes2*);
- área superficial do imóvel (*areaSuperficial_pes2*).

Esta última característica não foi levada em consideração nas próximas análises por já ser muito relacionada com a área habitável.

A fim de comparamos os imóveis, foi calculado a média de *preço/área habitável* e a média da qualidade da construção das cidades do Condado de King.

É possível notar nas figuras a seguir que os imóveis da cidade de Medida apresentam, em média, tanto a área de imóvel mais cara quanto a melhor qualidade de construção.


Média de preço/área das cidade do Condado de King
Woodinville
Vashon
Snoqualmie
Seattle
Sammamish
Renton
Redmond
North Bend
Mercer Island
Medina
Maple Valley
Kirkland
Kent
Kenmore
Issaquah
Federal Way
Fall City
Enumclaw
Duvall
Carnation
Bothell
ack Diamond
Bellevue
Auburn
0
100
200
300
400
500


Média da qualidade das construções nas cidades do Condado de King
Woodinville
Vashon
Snoqualmie
Seattle
Sammamish
Renton
Redmond
North Bend
Mercer Island
Medina
Maple Valley
Kirkland
Kent
Kenmore
Issaquah
Federal Way
Fall City
Enumclaw
Duvall
Carnation
Bothell
ack Diamond
Bellevue
Auburn
0
2
4
6
8
10

A fim de classificar os imóveis, uma nota foi para cada um deles.

Foram consideradas características desejadas nos imóveis para se realizar o investimento:

- *preço/área < média de preço/área da cidade* onde o imóvel está localizado. Desta forma, a relação *preço/área* do imóvel estará mais barata do que é cobrado em média naquela localidade.
- *qualidade da construcão > média da qualidade da construcão da cidade* onde imóvel está localizado. Desta forma, o imóvel terá uma qualidade superior a média da qualidade os imóveis naquela localidade.
- *área habitável > área habitável dos* 15 *vizinhos mais próximos*. Desta forma, o imóvel será maior que os imóveis vizinhos.

Cada uma dessas 3 relações será utilizada para calcular uma nota para os imóveis.

Como cada atributo possui uma correlação diferente com o preço do imóvel, foi criado um peso para cada uma das três notas.

Para isso, foi usado o valor de correlação entre preço e os atributos área habitável, qualidade da construção e área habitável dos 15 imóveis mais próximos.

As três notas atribuídas aos imóveis foram somadas e identificadas como *Nota_final* e os imóveis foram ordenados da maior para a menor nota final.

6 Sugestão dos imóveis

Os imóveis sugeridos para aquisição por parte da Roof Imóveis são os 5 melhores colocados no *ranking* criado a partir das notas. São eles:

id	**9421500130**
área habitável (pés2)	2760
preço (US$)	378.000,00
preço/área (US$/pés2)	136,95
cidade	Seattle
qualidade da construção	8
nota final	3,37

id	**1125079111**
área habitável (pés2)	6530
preço (US$)	1.600.000,00
preço/área (US$/pés2)	245,02
cidade	Carnation
qualidade da construção	11
nota final	2,72

id	**2722059275**
área habitável (pés2)	2290
preço (US$)	536.000,00
preço/área (US$/pés2)	234,06
cidade	Kent
qualidade da construção	7
nota final	2,67

id	7452500565
área habitável (pés2)	2710
preço (US$)	260.000,00
preço/área (US$/pés2)	95,94
cidade	Seattle
qualidade da construção	6
nota final	2,60

id	91756000025
área habitável (pés2)	7480
preço (US$)	800.000,00
preço/área (US$/pés2)	106,95
cidade	Seattle
qualidade da construção	11
nota final	2,55

Por fim, os imóveis que não são sugeridos para aquisição por parte da Roof Imóveis são os 5 piores colocados no *ranking* criado a partir das notas. São eles:

id	**3760500240**
área habitável (pés2)	750
preço (US$)	435.000,00
preço/área (US$/pés2)	580,00
cidade	Kirkland
qualidade da construção	4
nota final	0,45

id	**1222029077**
área habitável (pés2)	384
preço (US$)	265.000,00
preço/área (US$/pés2)	690,10
cidade	Vashon
qualidade da construção	4
nota final	0,42

id	**2420069251**
área habitável (pés2)	520
preço (US$)	262.000,00
preço/área (US$/pés2)	503,84
cidade	Enunclaw
qualidade da construção	3
nota final	0,42

id	**1925069006**
área habitável (pés2)	530
preço (US$)	355.000,00
preço/área (US$/pés2)	669,81
cidade	Sammamish
qualidade da construção	4
nota final	0,37

id	**3980300371**
área habitável (pés2)	290
preço (US$)	142.000,00
preço/área (US$/pés2)	489,65
cidade	Fall City
qualidade da construção	1
nota final	0,30