Projeto de Insights - Casas vendidas em King County (EUA), entre 2014/2015

Eric Brito, 2024

Sumário

1. Proposta do projeto

Projeto sugerido pelo cientista de dados sênior Meigarom Lopes, no seu artigo "5 projetos de Ciência de Dados para o seu portfólio de projetos", no portal Medium[1].

Perguntas a ser respondidas:

1. Quais casas o CEO da House Rocket deveria comprar e por qual preço de compra?

2. Uma vez a casa em posse da empresa, qual o melhor momento para vendê-las e qual seria o preço da venda?

3. A House Rocket deveria fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças? Qual o incremento no preço dado por cada opção de reforma?

Roteiro Sugerido para a Resolução:

1. Identifique a causa raíz. Porque o CEO fez essas perguntas? Se você fosse ele, porque você perguntaria isso? Quer aumentar receita? A empresa está indo bem? Anote essas causas.

2. Colete os dados.

3. Aplique uma limpeza nos dados. Entenda as variáveis disponíveis, possíveis valores faltantes, faça uma estatística descritiva para entender as características dos dados.

4. Levante Hipóteses sobre o Comportamento do Negócio. Casas com garagens são mais caras? Por quê?

5. Faça uma ótima Análise Exploratória de Dados. Quais hipóteses são falsas e quais são verdadeiras? Quais as correlações entre as variáveis e a variável resposta?

6. Escreve os Insights que você encontrou.

7. Escreva possíveis soluções para o problema do CEO.

Ferramentas da solução

Linguagem: Python.

IDE: VSCode (Jupyter Notebook).

Bibliotecas: Pandas, Plotly.

Código hospedado no Github[2].

Dataset (conjunto de dados)

House Sales in King County, USA[3].

Linhas: 21613; Colunas: 21.

---

[1] Disponível em: https://medium.com/comunidadeds/5-projetos-de-ci%C3%AAncia-de-dados-para-o-seu-portf%C3%B3lio-de-projetos-5fda646273e3.
[2] Disponível em: https://github.com/Mineirinhoua/Insights_casas_King_Count_EUA_14-15.
[3] Disponível em: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data.

2. Colunas do dataset e seus significados

Coluna	Descrição
id	ID exclusivo de cada venda.
date	Data da venda.
price	Preço da casa.
bedrooms	Número de quartos.
bathrooms	Número de banheiros. Completos ou parciais.
sqft_living	Área útil da casa, em pés quadrados.
sqft_lot	Tamanho do lote, em pés quadrados.
floors	Número de andares ou pavimentos da casa.
waterfront	Indicador binário, que é verdadeiro quando a casa está localizada à beira mar ou tem vista direta para a água.
view	Indicador de 0 a 4 sobre qualidade de vista da propriedade.
condition	Indicador de 1 a 5 sobre a condição da casa.
grade	Indicador de 1 a 13 sobre qualidade de construção e design.
sqft_above	Área das partes da casa acima do solo, em pés quadrados.
sqft_basement	Área do porão, em pés quadrados.
yr_built	Ano de construção da casa.
yr_renovated	Ano da última renovação/reforma da casa. Se ela nunca foi, valor é zero ou nulo.
zipcode	Código postal da localização da casa.
lat	Latitude da localização da casa.
long	Longitude da localização da casa.
sqft_living15	Área útil das 15 casas vizinhas mais próximas, em pés quadrados.
sqft_lot15	Tamanho do lote das 15 casas vizinhas mais próximas, em pés quadrados.

3. Objetivos e razões da análise.

Analisar o conjunto de dados relativos às vendas em King County, em Washiton nos EUA, entre maio de 2014 e de 2015. pode permitir a obtenção de valiosos insights, úteis para melhores decisões relacionadas à compra e venda de casas nesse condado estadunidense.

Entre as razões para o CEO elaborar as perguntas abordadas nessa análise, pode estar desde apenas a busca de lucros ainda maiores ou até uma situação mais crítica, como uma condição financeira delicada que exija decisões acertivas e melhor gerenciamento dos recursos da empresa. Qualquer que seja o caso, uma boa análise de dados pode auxiliar a observação e interpretação de padrões e tendências, que mostrem oportunidades e permitam a adoção de melhores estratégias de mercado para essa empresa.

4. Tratamento de dados

Antes das próximas etapas, é importante verificar problemas na base de dados e corrigi-los.

O primeiro passo foi verificar e resolver problemas gerais do conjunto de dados, como checar possíveis valores nulos. Esse conjunto de dados não possuia valores nulo, então não foi necessário apagá-los. Por outro lado, os ids da vendas, que deviam ser exclusivos, tinham algumas duplicatas, então foi necessário remover esses dados. Após isso, as colunas desnecessárias para a análise foram apagadas, no caso apenas a coluna id.

O segundo passo foi verificar e resolver problemas específicos de cada coluna. Essa etapa envolveu checar e corrigir os valores de colunas que devem estar dentro de uma faixa específica de valores, de fato estão. As datas precisaram ser convertidas de texto para o formato padrão das datas, o que permitiu checar se as datas realmente estavam dentro do tempo correto. Também foi preciso checar e corrigir os valores das colunas ilógicos, como por exemplo, uma casa com 800 banheiros ou -2 quartos. Até mesmo uma casa não muito grande com dezenas de quartos pode ser indicativo de problema.

Após realizar os ajustes, não se perdeu muitos dados, pois a qualidade dos dados originais era alta, com poucos problemas. É possível passar para a próxima etapa.

5. Hipóteses levantadas antes da análise

Hipóteses pré-análise:

1. Tamanho da área habitável da casa é fator que influencia muito o preço.

2. Tamanho do lote da casa é fator que influencia muito o preço.

3. Localização privilegiada influencia o aumento do preço.

4. Preços oscilam em curto período, de modo que no mesmo mês há momentos piores e melhores para realizar uma compra.

Hipóteses avaliadas pós-análise:

Hipótese	Análise	Justificativa
1	Verdadeira	O tamanho da área habitável é o dado que mais influencia o preço, conforme a análise da matriz de correlação.
2	Falsa	O tamanho do lote não tem tanta influência quanto esperado no preço, com base na análise da matriz de correlação.
3	Verdadeira	Casas cuja vizinhança possui um padrão de casas maiores, costumam ser mais caras. Isso não se deve inteiramente pela vizinhança, mas pelo fato da casa em questão também ser grande. Ainda assim, essa correlação deve ser considerada, como destacado na análise do mapa de calor.
4	Verdadeira	A análise mostrou que intervalos curtos de tempo, como as semanas de um mês são interessantes, pois de uma semana para outra há oscilações de preço médio consideráveis nas casas vendidas. Essa conclusão é amparada pelo gráfico de barras da figura 2.

6. Visualização e análise

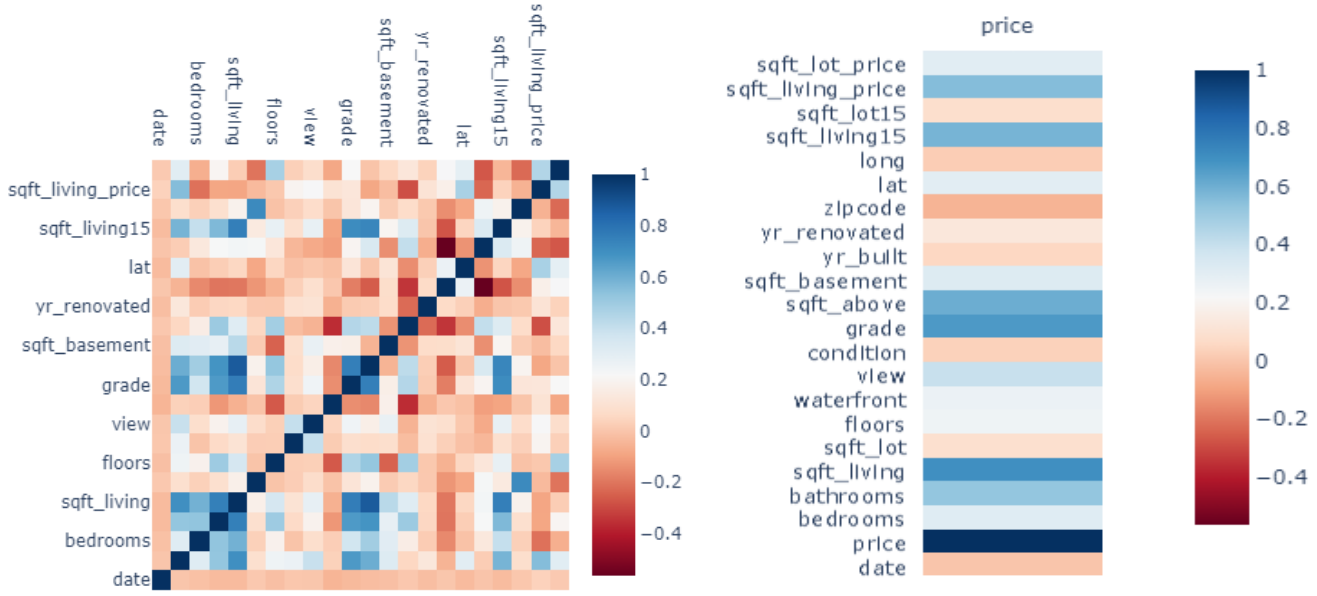
Para responder as perguntas do CEO, após limpar os dados, realizou-se uma análise inicial, que envolveu a geração de gráficos, que permite uma visualização melhor do problema.

6.1. Resposta da 1º pergunta

1. Quais casas o CEO da House Rocket deveria comprar e por qual preço de compra?

Um gráfico importante para apurar a correlações entre as variáveis e a varíavel resposta é o mapa de calor ou matriz de correlação. No gráfico a seguir, a variação de cores representa os valores de correlação entre as colunas do conjunto de dados. Tons mais puxados para o azul escuro indicam uma correlação de proporcionalidade, paralelismo, ou seja, o aumento ou diminuição de uma linha e coluna estão relacionados. Por outro lado, cores mais próximas do branco indicam uma relação de baixa correlação, enquanto cores mais avermelhadas indicam uma relação de proporcionalidade inversa, quando o valor de um deles aumenta ou diminui, o outro tende a seguir o inverso.

Figura 1 - Matriz de correlação



Segundo a matriz de correlação, os cinco fatores que mais se associam de forma linear ao preço são:

Fatores mais vinculados ao preço				
Área habitável	Nota	Acima	Arredores	Banheiros
0.7	0.66	0.6	0.58	0.52

Se esses fatores são importantes no preço de uma casa, é interessante também ver quais fatores estão associados aos principais deles, pois assim pode ser possível perceber um fator indireto que seja útil para desenvolver os fatores do preço.

Os cinco fatores mais associados à área habitável e à nota são respectivamente:

Fatores mais vinculados à àrea habitável				
Acima	Nota	Arredores	Banheiros	Preço
0.87	0.76	0.75	0.75	0.7

Fatores mais vinculados à nota				
Área habitável	Acima	Arredores	Preço	Banheiros
0.76	0.75	0.71	0.66	0.66

Percebe-se que não surgiram fatores indiretos, pois eles são os mesmos associados ao preço.

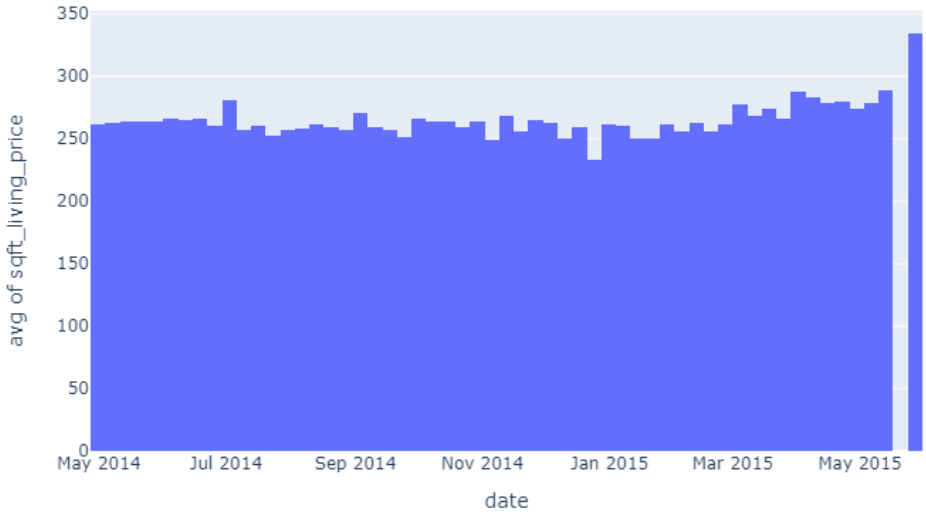
É importante não confundir a correlação linear com a causalidade, pois um valor geralmente aumentar ou dimunuir em paralelo a quando outro aumenta ou diminui, não implica que um seja a causa do outro. Por exemplo, encher uma casa de banheiros, visando aumentar o preço só porque a correlação apontou, não deve ser uma boa ideia, porque os dados acima mostram que é mais provável que a quantidade de banheiros cresça para suprir a demanda do tamanho da área habitável da casa, ao invés de ser um fator que vai fazer casas menores serem compradas. O fator "acima" também parece ser consequência da maior parte da área habitável das casas ser acima do porão, então ele não precisa ser focado. Os arredores mostram que há uma tendência de casas seguirem um padrão de tamanho relacionado a região, então casas maiores costumam estar em áreas com arredores e casas vizinhas maiores, o que também pode indicar regiões de maior ou menor padrão, dado que influencia na compra. Por fim, as notas só reforçam o efeito da área habitável, mas é uma coluna a se levar em consideração, pois não é só o tamanho de uma casa que indica sua qualidade.

Percebe-se nessa análise, como a área habitável é um fator importante, que puxa os outros com forte relação de causa e efeito, enquanto é pouco influenciado pelos outros, com baixa relação de causa e efeito. Por isso, ela é a coluna mais considerada no prosseguir da análise.

Como a área habitável é o fator que mais influencia o preço e as casas possuem tamanhos variados, mostrou-se proveitosa a criação da coluna "sqft_living_price", o preço do pé quadrado de área habitável. Seus valores são obtidos ao dividir o preço pela área habitável. Assim, obtem-se um valor relativo ao pé quadrado, que permite uma comporação mais justa entre casas de tamanhos distintos.

Através do seguinte gráfico de barras, é possível ver como o preço médio do pé quadrado de área habitável variou no tempo abrangido pelo dados.

Figura 2 - Gráfico de barras - Preço médio do pé de área habitável x Data



OBS: a parte vazia de Maio decorre de não haver dados desse curto período.

O período de menor preço médio relativo ao pé quadrado é a semana (21-27) de dezembro de 2014, com um valor de $232,74.

Uma estratégia alternativa/complementar

Dois dados interessantes que também podem ser levados em consideração na estratégia de compra, relacionados ao preço médio por pé quadrado são:

- **Ano de construção da casa**: como destacado pela cor avermelhada, geralmente é inversamente proporcional a esse valor, o que significa que casas mais antigas costumam ter o pé quadrado de área habitável mais caro;
- **Preço**: como enfatizado pela cor azulada, é razoavelmente proporcional a esse valor, o que indica que casas maiores, muitas vezes, também tem o preço médio por pé quadrado de área habitável maior.

Tendo esses dois dados em mente, procurar as casas que fogem dessas regras, pode ser a oportunidade de um bom negócio.

6.2. Resposta da 2º pergunta

2. Uma vez a casa em posse da empresa, qual o melhor momento para vendê-las e qual seria o preço da venda?

Por outro lado, o período em que o pé quadrado de área habitável é mais valorizado é na semana (24-30) de maio de 2015, com um valor de $334,92.

Então num cenário hipotético, considerando os valores relativos aos pés quadrados de área habitável, ao adquirir uma casa no melhor período para comprar, pelo valor médio de $232,74, e depois vendê-la no melhor período para vender, pelo valor médio de $334,92, obteria-se $102,17 de lucro, 43,9% do valor original.

6.3. Resposta da 3º pergunta

3. A House Rocket deveria fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças? Qual o incremento no preço dado por cada opção de reforma?

Sim, a empresa poderia reformar para aumentar o preço da venda. Casas com lote espaçoso são a oportunidade de expansão horizontal, para que com uma maior área habitável, a casa aumente de valor. Por outro lado, se não há esse espaço para explorar, é interessante expandir a casa verticalmente, especialmente se ela só tem um andar.

Com esse enfoque na área habitável não é preciso preocupar-se tanto com o número de garagens, banheiros ou quartos, porque o ênfase está no aumento de espaço utilizável. Porém, apesar deles impactarem menos no preço, ainda são diferenciais na escolha do cliente por uma casa específica, então é importante manter-se competitivo e atrativo. Um jeito simples de fazer isso é ser proporcional, por exemplo, não adicionar inúmeros banheiros numa casa pequena, adicionar conforme o tamanho delas e por consequência, de acordo com a provável necessidade.

Outro ponto a ser explorado nesse sentido, são os casos das casas com alguma ausência, por exemplo, sem banheiro ou garagem. Fazer essas adições, em casas que falta alguma coisa, para elas terem pelo menos uma, pode ser um atrativo importante para suas vendas. É preciso considerar alguns fatores, como se ela fica próximo de estações e tem menos nessecidade de garagem, ou se será apenas um escritório pequeno sem tanta necessidade de banheiro, ou ainda se o cliente pode preferir um quintal maior ao invés da reforma de expansão. Mas no geral, essas adições devem ser bem vindas.