Projeto Final

Contexto do Problema de Negócio

O investidor James Bauer gostaria de diversificar seus negócios e começar a investir em imóveis. Ele definiu que compraria imóveis na cidade de Nova York, nos Estados Unidos. Por ser um dos locais mais caros para se viver no País, ele acredita que obterá um retorno satisfatório de seus investimentos caso loque imóveis na cidade. Como todas as suas decisões são tomadas com base em dados, ele contratou você, cientista de dados, para ajudá-lo nessa empreitada.

James Bauer planeja inicialmente locar os imóveis adquiridos e por isso ele definiu que irá utilizar a plataforma Airbnb para esse fim. Para isso, ele lhe entregou uma base de dados públicos da empresa, contendo os dados do comportamento dos hosts e de seus imóveis.

Lembrando que o contexto, pessoas e perguntas são completamente fictícios e existem somente na minha imaginação.

O Desafio

James Bauer o contratou para realizar o estudo da base de dados fornecida e ajudá-lo com a escolha das regiões onde há maior locação e maiores preços, e que fiquem em regiões favoráveis da cidade de Nova York, pois ele acredita que essas características irão ajudá-lo a recuperar o dinheiro investido na aquisição desses imóveis mais rapidamente. Além disso, James Bauer pretende

Dessa forma, seu trabalho é realizar uma análise exploratória e responder às seguintes perguntas feitas pelo Sr. James Bauer:

1. Qual o `id` do imóvel com o aluguel (diária) mais caro da base de dados?
2. Qual o `id` do imóvel com o aluguel (diária) mais barato da base de dados?
3. Qual o `id` do imóvel que foi mais locado da base de dados?
4. Qual o `id` do imóvel que ficou mais tempo com o anúncio disponível, em dias, para locação na base de dados?

Hélio José da Silva Júnior - hjs.junior1@gmail.com - CPF: 016.073.125-90

5. Qual o `id` do imóvel que ficou menos tempo com o anúncio disponível, em dias, para locação na base de dados?
6. O imóvel com o maior valor de aluguel (diária) da base de dados é o imóvel que possui mais avaliações na base de dados?
7. O imóvel que possui a menor quantidade mínima de diárias para locação é também o imóvel que possui o aluguel mais caro?
8. Qual é a média do número mínimo de diárias para locação de um imóvel?
9. Qual é o `id` do imóvel com a menor quantidade mínima de diárias para locação da base de dados?
10. Qual é o `id` do host que possui o imóvel mais alugado na base de dados?
11. Qual é o `id` do host que possui o imóvel menos alugado na base de dados?
12. Qual é o `id` do host que possui o imóvel com mais avaliações na base de dados?
13. Qual é o `id` do host que possui a maior quantidade de imóveis cadastrados na base de dados?
14. Qual o `id` do host que possui o imóvel com a última avaliação feita na base de dados?
15. Qual o `id` do host que possui mais imóveis ativos dentro da base de dados?
16. O host que possui mais imóveis é o host que também possui mais avaliações?
17. Qual é a categoria que mais possui imóveis dentro da base de dados?
18. Qual é a categoria que menos possui imóveis dentro da base de dados?
19. A região de Manhattan é a região que mais possui imóveis ativos para locação, da categoria `Private room`?
20. Qual a categoria de imóvel que possui a maior média de tempo de disponibilidade para serem locados da região do `Bronx`? Considere somente imóveis ativos
21. Qual a categoria de imóvel ativo que possui o maior valor de aluguel (diária) na região de `Manhattan`?
22. A categoria de imóvel `Private Room` é a categoria que fica mais tempo disponível para locação?
23. Qual a categoria de imóvel que fica mais tempo disponível, na média, para locação?

Hélio José da Silva Júnior - hjs.junior1@gmail.com - CPF: 016.073.125-90

24. A categoria de imóvel que fica menos tempo disponível, na média, para locação é a categoria que possui, em média, o menor aluguel (diária)?
25. A categoria de imóvel `Entire home/apt` é a categoria que possui, na média, o maior valor de aluguel?
26. A categoria de imóvel `Entire home/apt` é a categoria que possui, na média, menos locações?
27. A categoria de imóvel `Private Room` na região de `Manhattan`, na média, é a categoria que possui o menor valor (diária) de locação, comparado as outras categorias na mesma região?
28. Qual a região que possui a maior quantidade de imóveis?
29. Qual a região que possui a menor quantidade de imóveis?
30. A região que possui a maior quantidade de imóveis é também a região que possui os imóveis mais locados?
31. Qual a região que possui a menor quantidade de imóveis locados dentro da base de dados?
32. A região de `Manhattan` é a região que possui, na média, os maiores aluguéis (diárias) dentro da base de dados?
33. A região de `Queens` é a região que possui, na média, os menores aluguéis (diárias) dentro da base de dados?
34. Qual a região que possui, na média, os imóveis com os menores aluguéis dentro da base de dados?
35. Qual a região que possui os imóveis que ficam, na média, menos tempo disponíveis para aluguel? Ou seja, são alugados mais rápidos na média?
36. Qual o bairro possui a maior quantidade de imóveis ativos para locação?
37. Qual o bairro possui a menor quantidade de imóveis ativos para locação?
38. Qual bairro está o imóvel com o maior valor de aluguel?
39. Qual bairro está o imóvel com o menor valor de aluguel?
40. Qual região que possui o bairro com mais imóveis ativos disponíveis para locação?
41. Qual região que possui o bairro com menos imóveis ativos disponíveis para locação?
42. O Bairro `Upper West Side`, na região de `Manhattan`, é o bairro que possui, na média, o maior aluguel dentro da base de dados?
43. Qual é o bairro que possui, na média, o aluguel mais caro da base de dados?

44. Qual o id do imóvel que possui a melhor rentabilidade da base de dados? Levando em consideração o aluguel mais caro e imóvel mais rápido de alugar? Utilize somente os imóveis que estejam ativos e que possuam pelo menos uma avaliação. Utilize a seguinte fórmula para verificar qual o melhor imóvel:

$$rentabilidade = \frac{price * (minimum\_nights + 1) * number\_of\_reviews}{\sqrt{availability\_365}}$$

45. Qual a região em que fica o melhor bairro para se adquirir um imóvel, visando a melhor rentabilidade média? Utilize como métrica o índice criado na questão anterior.
46. Qual a região em que fica o pior bairro para se adquirir um imóvel, visando a melhor rentabilidade média? Utilize como métrica o índice criado na questão 44.
47. Levando em conta o bairro encontrado na questão anterior, verifique qual é o `id` do imóvel ativo que possui a pior rentabilidade desse bairro
48. Levando em conta o bairro da questão 45, verifique qual o `id` do imóvel ativo que possui a melhor rentabilidade desse bairro
49. Caso eu, investidor, compre o imóvel da questão anterior com um investimento de U$ 1.000.000,00, quantas vezes eu teria que locá-lo para ter o retorno do investimento feito? Utilize a fórmula abaixo para calcular o tempo de retorno do investimento

$$return\_investment = \frac{investment}{price * (minimum\_nights + 1)}$$

50. Se eu desejasse comprar um imóvel que esteja ativo e com pelo menos uma avaliação em um dos bairros abaixo, qual seria o id do imóvel com a melhor rentabilidade dentre esses bairros? Utilize o índice calculado na questão 44.
    - East Harlem
    - Harlem
    - Midtown
    - Morningside Heights
    - Upper West Side
    - Upper East Side

Para responder às perguntas, que contenham mais de um imóvel como resposta, selecione sempre o imóvel que for mais antigo no Dataset. Ou seja, o imóvel que possui a coluna “id” menor.

Os Dados

O conjunto de dados que representam o contexto está disponível na plataforma do Kaggle. O link para acesso aos dados :

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Como Solucionar Esse Desafio

Aqui vão algumas dicas para você começar a resolver esse problema:

- **Tenha calma e não tenha medo:** Crie suposições e faça testes, dando um passo de cada vez para, a cada novo passo, estar mais próximo da resposta.
- **Responda as perguntas antes de Codificar:** Como você faria, se não tivesse que programar, para responder as perguntas feitas? Pensando antes em como responder, facilitará no momento de codificar o algoritmo para achar a resposta, porque isso faz parte do planejamento da solução!
- **Tenha Paciência e Resiliência:** Criar soluções e, principalmente, pensar e planejá-las leva tempo. Assuma uma postura resiliente e não desista! Afinal, você quer se tornar um Cientista de Dados, ganhar ótimos salários e trabalhar em ótimas empresas, certo?
- **A Comunidade:** Caso você tenha tentado várias estratégias e não tenha chegado a uma solução ou não tenha avançado, peça ajuda dentro da Comunidade de DS! Estamos todos para nos ajudar nessa jornada que é a Ciência de Dados.

Roteiro Sugerido para Resolução

Um roteiro que pode ser usado como resolução do desafio, seria:

1. Entenda o problema de negócio e suas perguntas:

    a. Por que o investidor fez essas perguntas?
    b. Se você fosse ele, por que essas informações seriam úteis para você?
    c. Anote as suas respostas e possíveis causas para as perguntas terem sido feitas.
2. Colete os dados:
    a. Os dados estão disponíveis no link neste post
3. Faça uma limpeza nos dados:
    a. Entenda as variáveis disponíveis na base de dados fazendo uma tabela de estatística descritiva.
    b. Verifique se há dados faltantes e quais estratégias poderiam ser utilizadas para preenchê-los.
4. Levante hipóteses sobre o comportamento do Negócio:
    a. Por exemplo:
        i. Imóveis que são alugados por inteiro possuem o valor do aluguel mais caro?
        ii. A região de “Manhattan” possui, na média, aluguéis mais caros que as outras regiões?
    b. Valide as suas hipóteses com testes estatísticos e gráficos.
5. Explore os dados e responda as perguntas:
    a. Comece respondendo as perguntas sem usar programação, para planejar a sua solução. Ou seja, responda primeiro como você faria para responder a pergunta. Por exemplo: Calcularia a média dos valores dos aluguéis dos imóveis de uma determinada região de um determinado bairro.
    b. Com o planejamento feito, implemente os passos/ações utilizando a linguagem de programação escolhida.
    c. Se necessário, utilize gráficos para consolidar e validar a sua resposta.
    d. Caso uma pergunta específica se mostre falsa, encontre a situação ou fato que a torne falsa. Por exemplo, os imóveis da região de “Manhattan” possuem os aluguéis mais baratos? Caso essa pergunta seja falsa, verifique qual a região que possui os aluguéis mais baratos.
6. Deixe suas respostas disponíveis:
    a. Pense e planeje uma solução para deixar os seus Insights e as respostas disponíveis para o investidor possa acessá-las

Hélio José da Silva Júnior - hjs.junior1@gmail.com - CPF: 016.073.125-90

O Ferramental da Solução

Como ideia desse projeto é colocar em prática e revisar todas as técnicas e conhecimento adquirido dentro do curso de Python do Zero ao DS, recomendo a utilização das ferramentas usadas durante o curso, como o Jupyter Notebook, para prototipar a solução, um editor de texto ou IDE para te auxiliar no desenvolvimento da solução que será colocada em produção e o framework Streamlit para possibilitar a criação de Web Apps.

Porém, caso seja de seu desejo, utilize as ferramentas que forem do seu agrado e que te tornem mais produtivo! A melhor ferramenta não é a que vemos os outros usarem, e sim a que nos torna mais produtivos e eficientes ao realizar o nosso trabalho!

Aproveite para treinar e revisar todo o conteúdo visto, melhorando assim a sua velocidade de manipulação de dados, raciocínio analítico e resolução de problemas! Tenho certeza de que com esses projetos e tempo, você escreverá códigos e ferramentas cada vez mais robustas e ficará cada vez mais fluente e seguro nas técnicas e ferramentas usadas dentro do universo de Ciência de Dados.

Vá em frente!

Como sempre conversamos dentro da Comunidade, não existe caminho fácil e de curto prazo em nenhuma profissão, e ciência de dados não é exceção. Através de estudo e desenvolvimento de projeto, você irá aumentar a sua capacidade de resolver problemas e entregar resultados.

E os seus projetos de portfólio servem para esse fim: Demonstrar as suas capacidades de resolver problemas. Ou seja, demonstram que você é tão capaz quanto um cientista de dados que já atua profissionalmente nas empresas.

Conclusão

Nesse post, você recebeu um desafio de Ciência de Dados muito próximo dos desafios reais das empresas.

Hélio José da Silva Júnior - hjs.junior1@gmail.com - CPF: 016.073.125-90

Os problemas chegam em forma de perguntas abertas, desestruturadas e sem nenhuma dica sobre como resolver.

É papel do Cientista de Dados entender a causa raiz, planejar o desenvolvimento e criar a melhor solução para o problema de negócio.

Não se esqueça de acompanhar o canal “Seja um Data Scientist” e o Instagram @meigarom para mais conteúdos. Caso você tenha LinkedIn, não se esqueça de conectar comigo, é só procurar por Meigarom, está fácil de encontrar!!

Bons estudos!!

Hélio José da Silva Júnior - hjs.junior1@gmail.com - CPF: 016.073.125-90