Teste Técnico - Analista de Dados

Este teste tem como objetivo avaliar sua habilidade prática em SQL e Python.

📂 Parte 1 - SQL (BigQuery)

Você recebeu acesso de leitura ao dataset do teste no BigQuery (`karhub-techtest.ecomm`) através do e-mail que informou no processo seletivo.
Para acessar:

1. Entre em Google BigQuery Console: https://console.cloud.google.com/bigquery
2. Certifique-se de estar logado com o mesmo e-mail que enviou para nosso time.
3. No painel esquerdo, navegue até: **`karhub-techtest`** → **dataset `ecomm`**
4. Lá você encontrará as tabelas necessárias para responder às questões do teste.

Tabelas disponíveis:

Tabela	Descrição	Campos
products	Catálogo de produtos	`product_id` (INT), `product_name` (STRING), `category` (STRING), `base_price` (NUMERIC), `short_name` (STRING)
customers	Dados dos clientes	`customer_id` (INT), `customer_name` (STRING), `region` (STRING), `sign_up_date` (DATE)
orders	Pedidos realizados	`order_id` (INT), `customer_id` (INT), `order_date` (DATE), `total_amount` (NUMERIC), `order_ts` (TIMESTAMP), `order_datetime_str` (STRING)
order_items	Itens dentro dos pedidos	`order_id` (INT), `product_id` (INT), `quantity` (INT),

		`unit_price` (NUMERIC), `line_amount` (NUMERIC)

Observações:

1. A tabela `orders` contém **duplicatas** (mesmo `order_id` com múltiplas versões). Sempre que usar `orders`, **deduplique** mantendo a versão mais recente (maior `order_ts`)
2. `orders.total_amount` é a soma dos itens (`line_amount`) da tabela `order_items`.
3. A tabela `customers` contém nomes **desnormalizados** (variações de maiúsculas/minúsculas e espaços).
4. Ao exibir nomes de produtos, dê preferência para a coluna short_name e, se não tiver, use a product_name.

Com **SQL** responda as questões abaixo:

1. Liste os **10 produtos mais vendidos** em número de itens no ano de 2025.
2. Liste os **5 clientes com maior gasto total** no ano de 2025.
3. Quais produtos você sugere que recebam maior investimento em divulgação? Responda com a query e com a justificativa para seleção desses produtos.
4. Considerando a região dos clientes, quais insights sobre logística você pode gerar?
5. Quais clientes deveriam ser incluídos em uma campanha de reativação?
6. Analisando a aquisição de novos clientes com base na data de cadastro, qual mês se destacou, indicando a possível eficácia de iniciativas de marketing?

📂 Parte 2 - Python

Você receberá um arquivo chamado **`sales.csv`**, contendo colunas:

Coluna	Descrição
`order_id`	Identificador do pedido
`customer_name`	Nome do cliente (pode variar em maiúsculo/minúsculo e conter espaços no final)
`product_name`	Nome do produto (idem acima, pode variar)
`quantity`	Quantidade de itens no pedido
`price`	Preço unitário do produto

`order_date`	Data do pedido
`region`	Região do cliente
`revenue`	Receita do item (quantity × price)
`discount`	Desconto aplicado (valor entre 0 e 1, ex.: 0.10 = 10%)

*Observação: como os nomes podem variar em formatação, considere se será necessário **normalizar os textos** antes de analisar.*

Com os dados desse arquivo, retorne uma planilha/arquivo csv com os nomes das colunas `customer_name` e `product_name` normalizados e com uma nova coluna considerando o valor com desconto de cada linha utilizando as colunas revenue e discount.

O que esperamos:

A entrega pode ser feita em um arquivo .zip contendo os arquivos abaixo, ou um repositório no git.

- Parte 1 (SQL): entregue as *queries* que você construiu em um documento (.sql, .docx ou .txt)
- Parte 2 (Python): entregue um notebook (.ipynb) com suas respostas.