Case Técnico

Programa de Estágio em Data Science 2019

Olá, candidato. Parabéns por ter chegado até aqui!

Este case técnico foi desenvolvido para conhecermos suas habilidades e familiaridade com assuntos que você encontrará no seu dia-a-dia como estagiário no Jeitto. Os tópicos abordados são:

- Exploração de dados
- Visualização e apresentação de dados
- Conhecimentos sobre bancos de dados
- Manipulação de dados com SQL e Python (Pandas)

A prova foi pensada para ser relativamente longa e exercitar sua capacidade de buscar informações na Internet. Não queremos que você saiba tudo de cabeça, mas que saiba buscar ajuda se encontrar dificuldade em algum ponto. O principal objetivo é entendermos a forma como você ataca um problema, organiza seu raciocínio em torno dele e apresenta os resultados. A interpretação das questões é parte do processo e algumas questões são propositalmente abertas à interpretação, mas caso haja algum enunciado duvidoso, o autor do case fica à disposição para esclarecer quaisquer dúvidas no e-mail pablo.brenner@jeitto.com.br.

Formato da entrega

Ao final das 24h disponíveis para a entrega do case enviar para o e-mail pablo.brenner@jeitto.com.br um relatório contendo a resposta para cada questão, assim como o código que acompanha algumas questões e, se necessário, uma breve descrição de como você buscou a solução. Utilize sua criatividade para apresentar os resultados e entregar o seu trabalho o que pode ser feito em PowerPoint, em um arquivo PDF ou até mesmo em um arquivo do Jupyter Notebook.

Critérios da avaliação

Queremos entender a forma como você atacou o problema, assim como a qualidade do código e a linha de raciocínio estabelecida. Caso você não tenha conseguido resolver alguma questão, não há problema! Descreva com detalhes o seu esforço de resolução e onde ficou bloqueado.

Atenção: a consulta a quaisquer materiais é livre. A cópia deles, não. Sabemos identificar a consistência da qualidade de código e escrita entre as questões e vamos identificar quais questões foram copiadas.

Questão 1: Exploração e Visualização de dados

O CEO da sua empresa deseja entender melhor o perfil e comportamentos da base de clientes e por isso acionou você como profissional de dados para gerar insights que possam ser interessantes e valiosos para o negócio. Para tal tarefa você dispõe da base de dados **base_clientes.xlsx** para realizar o trabalho exploratório de dados e apresentar seus achados ao CEO com possíveis recomendações.

Desenvolva uma análise exploratória com visualizações de dados e descreva insights que você vai achar a partir dos dados. Estruture a sua entrega como se estivesse apresentando os resultados em uma reunião de lideranças do negócio, utilizando sua criatividade e capacidade de storytelling para comunicar efetivamente através do seu material os seus achados.

Dica: essa é a questão com maior peso na avaliação, dedique uma proporção superior a sua solução em relação as demais.

Questão 2: Conceitos sobre Bancos de dados

Atualmente existem diversos tipos de bancos de dados disponíveis no mercado, muitos deles open source e com características de funcionamento e aplicação diferentes entre si. Descreva as principais características e diferenças entre os seguintes tipos de bancos de dados:

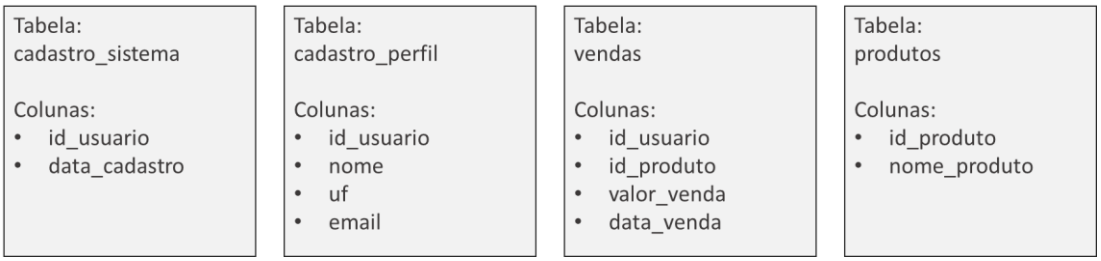
• Bancos de dados relacionais

• Bancos de dados não relacionais

• Bancos de dados de grafos

Questão 3: Manipulação de dados com SQL

Você possui acesso as tabelas de um banco relacional conforme o esquema abaixo de colunas em cada tabela (cada caixa é uma tabela com os nomes das colunas disponíveis).



Construa as consultas SQL para responder as seguintes questões (você pode utilizar a sintaxe SQL que preferir):

• Gere uma relação dos clientes da empresa com nome, uf e e-mail ordenados do mais recente para o mais antigo

• Construa uma consulta que retorne os produtos mais vendidos com nome do produto e quantidade de vendas

• Agrupe a consulta anterior por meses do ano

Questão 4: Conceitos de Engenharia de Dados

Para que as consultas que você construiu na questão anteriores sejam executadas com performance você precisará que as tabelas estejam corretamente indexadas. Considerando isso responda as seguintes questões:

• O que é o índice em uma tabela de banco de dados relacional? Qual a sua necessidade?

• Quais índices deveriam ser criados para acelerar as consultas da questão anterior?

• Qual o código SQL em MySQL ou PostgreSQL para a criação do índice na coluna id_produto da tabela produtos?

Questão 5: Manipulação de dados com Python (Pandas)

Nessa questão você deverá responder as questões abaixo utilizando a base base_clientes.xlsx e os recursos de manipulação de dados da biblioteca Pandas. Para esta questão você deverá enviar o código Python que escreveu para chegar em cada um dos resultados bem como o dataframe resultante.

• Qual o total de vendas de cada produto por estado?

• Qual o total de vendas, total de valor vendido e ticket médio de cada produto?

• Quantos clientes existem na base com as faixas etárias de 0-18, 19-25, 26-32, 32-45 e 45+ anos?

• Qual é o principal produto da primeira compra conforme as faixas etárias da questão anterior?