Teste Técnico

Posição de Data Analyst Jr - Buraco Jogatina

Seja bem-vindo(a) nesta etapa do processo e agradecemos a sua participação.

Antes de começar, queremos dar algumas dicas

- O teste é composto por 3 componentes principais: *SQL*, *Python* (raciocínio lógico e matemático) e *Conceitos de estatística e Machine Learning*
- Cada componente apresenta questões de diferente tipo, portanto, avalie e entenda-as com paciência e calma. Você contará com um prazo mais do que o suficiente para completar o teste.
- **E o mais importante:** Algumas perguntas efetivamente foram desenhadas para questionar sobre conhecimentos técnicos que podem se mostrar mais avançados do que o escopo da posição. Portanto, se deixa de respondê-las, não fique preocupado(a)!

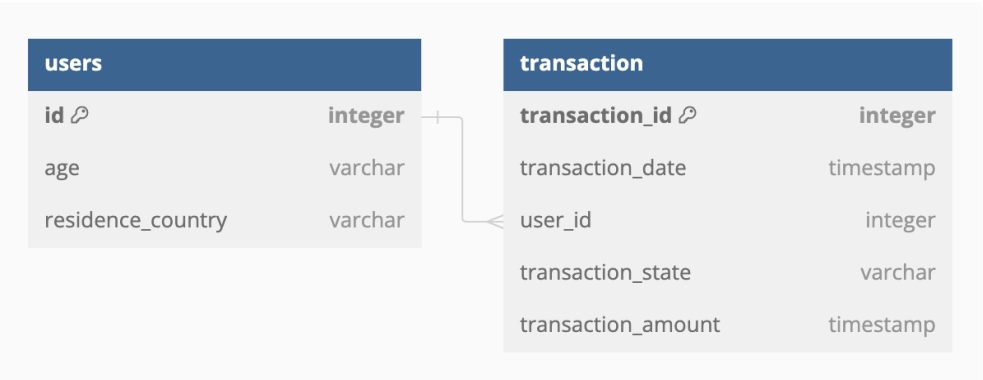
Novamente, sucesso e boa sorte.

Tarefa 01: SQL

Um sistema de pagamentos, registra as transações dos seus usuários em em duas tabelas:

- **Users:** Contém o registro de usuários do sistema. Os seus atributos são
  - **id:** Identificador único de usuário
  - **age (idade):** em anos
  - **país de residência:** país onde o usuários indicou que reside

- **Transactions:** Que contém o registro de todas as transações que os usuários realizam através do sistema.
  - **transaction_id:** Identificador único da transação
  - **transaction_date:** Data na qual a transação foi realizada. O tipo é timestamp, ou seja, é um tipo de data que contém o ano, mês, dia, hora, minuto e segundos.
  - **user_id**: Identificador do usuário que realizou a transação (é uma chave externa à tabela users)
  - **transaction_state:** Campo que indica o estado da transação que pode ser INITIATED, SUCCESS, FRAUD or CANCELLED.
  - **transaction_amount:** Valor da transação em USD



Com estes dados disponíveis, precisamos gerar queries SQL para responder às seguintes perguntas:

A. Qual é a idade média de usuários do sistema por país
B. Qual é o país com a maior quantidade de dinheiro transacionado (considere só transações finalizadas com sucesso ou "SUCCESS")
C. Qual é o país com maior taxa de fraude em porcentagem respeito ao número de transações totais no país
D. Na mesma linha da pergunta anterior, responda qual é a faixa de idade de usuários que mais cometem fraude (em percentagem). Separe as faixas etárias em: (*<18 anos, 18-30 anos, 30 - 45 anos, 45 - 60 anos, 60 > anos)* e considere que para responder essa pergunta, você deverá considerar o fato que um usuário pode ter

executado várias transações, das quais poucas (ou muitas) podem ter sido fraude entre as demais.

E. Imagine que a camada executiva da empresa dona do sistema, precisa criar um Dashboard para monitorar o estado das transações nos últimos 3 dias. Para isso você precisa criar uma query SQL que calcule o número e dinheiro das transações não finalizadas, número e dinheiro de transações finalizadas com sucesso (SUCCESS), o número e dinheiro de transações canceladas (CANCELLED), o número e dinheiro de fraudes (FRAUD), **agrupado por país**, **nos 3 dias anteriores de quando o executivo da empresa consulte seu Dashboard**.

Tarefa 02: Python

Numa tribo ancestral e muito desenvolvida, orientada principalmente por uma cultura lógica e matemática, os nativos estão interessados no estudo das linguagens de outras tribos e civilizações.

Durante seus estudos, estes perceberam que as linguagens dessas outras civilizações são compostas por um conjunto de símbolos (letras do alfabeto) que são agrupados e combinados para representar conceitos (palavras). Assim mesmo, estes também sabem que cada civilização possui um alfabeto específico.

Na tentativa da tribo de avançar nos seus estudos, **estes desejam saber qual seria o número de palavras possíveis a serem criadas em função do tamanho da palavra e o conjunto de símbolos (alfabeto) da civilização em estudo**, independentemente se essas "palavras" representam algum significado ou não.

A) Você, que compartilha os atributos da tribo enquanto as capacidades analíticas e lógicas, **precisa ajudá-la escrevendo um algoritmo para fazer o cálculo descrito acima**, considerando que tribo oferecerá para você o alfabeto da civilização e o tamanho da palavra/combinação.

   Exemplo: Se a tribo deseja avaliar o número de palavras possíveis juntando 2 símbolos considerando um alfabeto composto pelos seguintes 4 símbolos: {@,d,2,b} (exemplos de palavras possíveis abaixo)

   - "@@"
   - "@d"
   - …
   - "dd"
   - "d2"
   - …
   - "2b"
   - "b2"
   - …
   - "bb"

   Seu algoritmo deveria ser capaz de calcular que o número total de palavras possíveis (independente se tem significado ou não) **é de 16**

   OBS: NÃO PRECISA ESCREVER A LISTA DAS PALAVRAS, A TRIBO SÓ PRECISA DO NÚMERO!

B) Para ajudar a tribo ainda mais! Você deverá modificar seu algoritmo (ou talvez escrever um novo) para fazer o mesmo cálculo, só que agora, as palavras não podem ter símbolos repetidos.

Exemplo:

Do anterior exemplo**,** palavras como "@@" ou "dd", não devem ser contadas.

Para uma palavra de tamanho 3 usando o mesmo alfabeto ({@,d,2,b}), palavras "222", "@d@" ou "2dd" não deveriam ser contadas. Palavras como "@d2" ou "b@2", devem ser contadas.

Neste sentido, o (novo/modificado) algoritmo, para uma palavra de tamanho 2 usando o mesmo alfabeto (de 4 símbolos), **deverá ser capaz de calcular que a saída é 12**.

DICA: Pode ajudar, usar a função factorial.

Tarefa 03: Estatística e Machine Learning

1. Dado um dataset das estaturas (em centímetros) de 13 indivíduos:

   Dataset = {175, 166, 183, 193, 155, 177, 173, 171, 162, 185, 176, 161, 188}

   Calcule a média, a mediana, 20 percentil, 80 percentil e desvio padrão. Considerando os dados, responda a que distribuição (paramétrica) os dados se aproximam.

2. Com suas próprias palavras, explique em que consiste o teorema central do limite, e se possível, mencione a sua importância no campo da inferência estatística.

3. Suponha que numa escola, 2 grupos diferentes de estudantes (Grupo A e Grupo B) fazem o mesmo teste de matemáticas. As pontuações para cada grupo são dados pelos seguintes datasets:

   Grupo A: {80, 85, 88, 90, 92, 75, 78}
   Grupo B: {75, 78, 82, 85, 87, 93, 99}

   Elabore um teste de hipótese para determinar se existe uma diferença estatisticamente significativa entre a média das pontuações dos dois grupos com uma confiança de 95%.

4. Determine a hipótese nula (Ho) e hipótese alternativa (Ha) do seguinte cenário: Uma empresa afirma que o tempo médio dos seus produtos é menos de 4 dias. Você, conta com uma amostra dessas entregas para validar estadísticamente essa afirmação.

5. Calcule o coeficiente de correlação de pearson das variáveis "*Número de aparições Zendaya em filmes por ano*" e "*Número de pessoas afogadas em piscinas no Brasil ao ano*"

Ano	Número de aparições Zendaya em filmes por ano	Número de pessoas afogadas em piscinas no *Brasil ao* ano
2015	4	107
2016	7	146
2017	16	178
2018	21	199

2019	26	221
2020	2	114
2021	6	133
2022	12	159
2023	16	183
2024	32	215

Em base nos seus resultados, considera que a variável "*Número de aparições Zendaya em filmes por ano*" é um bom preditor do número de pessoas afogadas no Brasil?

Justifique a sua resposta.

6. Explique a diferença entre amostragem **estratificada** e amostragem **randômica** ou aleatória. Discuta quais são as vantagens e desvantagens de cada uma e de exemplos de casos onde uma abordagem é mais adequada que a outra.

7. Se você treina um modelo de Machine Learning (ou estadístico), como você identificaria se seu modelo tem uma alta variância (overfitting) ou um alto viés (bias, ou underfitting). Caso seu modelo apresenta alta variância, como você resolveria esse problema?

8. Considere o seguinte modelo de regressão:

$$Salário = 1200 + 500.Idade + 600.\ Têm\ Faculdade + 50.Têm\ Linkedin$$

   A. Interprete o efeito da Idade na variável salário
   B. Que acontece com as pessoas que não tem faculdade?
   C. Considere agora que o desvio padrão do **coeficiente** da variável "$Têm\ Linkedin$" é de 47, por tanto o seu p-valor é ~0.92 (como visto na equação abaixo). O que isso implica para o modelo em questão? Essa variável é relevante?

$Salário =$	$1200 +$	$500.Idade +$	$600.\ Têm\ Faculdade +$	$50.Têm\ Linkedin$
std	(200)	(125)	(150)	(47)
p-value	0.01	0.023	0.032	0.92