



Processo Seletivo: Cientista de Dados (Analista Jr.)

Este case é constituído por 2 exercícios. No primeiro, você deverá trabalhar no tratamento, transformação e análise dos dados contidos em arquivo, assim como sua capacidade de relatar informações e insights. Já no segundo exercício, você deverá mostrar sua habilidade na elaboração de queries em bancos de dados baseados em SQL.

Para a resolução do exercício 1, temos preferência por Python. É importante descrever detalhadamente todos os recursos utilizados para a execução de seu código. Caso ele não seja executado por algum motivo ou erro, ele será desconsiderado em sua avaliação.

Fique atento(a)! Em sua entrega deve conter o código utilizado para tratar/modelar dados e o relatório com todas argumentações e insights do exercício 1, além das queries do exercício 2.

FAÇA UM ÓTIMO TESTE!

Exercício 1: Análise de Internações no Sistema de Saúde Brasileiro

Você foi contratado(a) para fazer uma análise apurada do número de internações no sistema de saúde brasileiro. Esta análise é de extrema importância para tomada de decisões que deverão contribuir para melhorias no sistema e planejamento estratégico.

Os dados em anexo (case_internacao_SUS.xls) são referentes às internações que ocorreram no país durante o período de dezembro de 2017 a julho de 2019, separados por região e unidade de federação (Fonte: Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS)).

Acesse <http://tabnet.datasus.gov.br/cgi/sih/sxdescr.htm> para obter uma descrição detalhada dos dados.

Desafios:

1. Tratamento dos dados

- 1.1. Muitas vezes, cerca de 70% do tempo de um projeto é despendido na coleta e tratamento dos dados. Sabendo disso, leia o arquivo e o transforme de modo a ter mais facilidade em analisar os dados. Lembre-se que essa etapa poderá te dar bons insumos. Portanto, capriche!

2. Análise

- 2.1. Dados tratados, bora explorá-los? Faça uma boa EDA e não esqueça de anotar todos os insights que você obter. Gráficos e informações sem uma boa interpretação não valem, ok?

3. Modelagem

- 3.1. Agora que já tem certa intimidade com os dados, cite pelo menos 2 métodos possíveis para estimar os dados para os meses faltantes. Tente não se complicar aqui. Utilize os métodos mais simples e mais funcionais possíveis. Neste tópico, é importante que argumente o porquê dos métodos recomendados.
- 3.2. Escolha um desses métodos e estime a) o número de Internações e b) o Valor Total das internações nos períodos faltantes.

3.3. Crie um modelo que preveja a) as Internações, b) o número de Óbitos e c) o Valor Médio de AIH pelos próximos 6 meses. Explique a escolha do modelo e quais parâmetros utilizou para serem input no modelo.

4. Planejamento estratégico

4.1. Com base nos dados e nas suas análises, que tipo de estratégia você sugeriria para diminuir o número de internações em hospitais do SUS? E para o Estado de São Paulo? Quais especificidades deveriam ser levadas em conta?

Ao final, escreva um relatório contendo todos os gráficos gerados bem como as argumentações, análises e insights de sua EDA (item 2.1). Neste relatório também deverá conter a argumentação do porquê dos métodos citados no item 3.1 são cabíveis aos dados disponibilizados. Além disso, devem ser incorporadas ao relatório as análises ou argumentos que te levaram a optar pelo método escolhido para a modelagem requerida nos itens 3.2 e 3.3. Não esqueça de anexar ao relatório toda argumentação e apontamentos para o questionamento feito no item 4.1.

Dica 1: O relatório pode ser entregue em qualquer formato (docx, ppt, docs, Jupyter Notebook, etc). Tenha em mente que o documento gerado deve ser de fácil acesso e uso mais simplificado possível. Ele poderá ser usado por diferentes profissionais com as mais variadas skills. Portanto, seja criativo e faça sua argumentação ser mais clara possível. Faça um bom storytelling!

Dica 2: Não esqueça de inserir comentários em seu código. É importante para entender seu racional.

Exercício 2: Queries em SQL

Imagine uma base de dados de uma universidade que contenha as seguintes tabelas e seus respectivos campos organizados da seguinte forma:

tabela_alunos → colunas = (ra, nome, endereco, cidade);

tabela_disciplinas → colunas = (cod_disc, nome_disc, carga_hor)

tabela_professores → colunas = (cod_prof, nome, endereco, cidade)

tabela_turma → colunas = (cod_disc, cod_turma, cod_prof, ano, semestre)

tabela_historico, colunas = (ra, cod_disc, cod_turma, cod_prof, ano, semestre, frequencia, nota)

- a) Monte uma consulta que informe quantos alunos cursaram a disciplina de Banco de Dados em 2019 e 2020;
- b) Monte uma consulta que exiba os nomes dos alunos que não foram reprovados em nenhuma disciplina (considere nota mínima igual a 6 para aprovação).

Dicionário de colunas

\$ ra	: registro acadêmico do aluno
\$ nome	
\$ endereco	
\$ cidade	
\$ cod_disc	: código da disciplina
\$ nome_disc	: nome da disciplina
\$ carga_hor	: carga horária
\$ cod_prof	: código do professor
\$ cod_turma	: código da turma
\$ ano	
\$ semestre	
\$ frequencia	: quantidade de presenças no período letivo
\$ nota	: nota obtida na disciplina