

Aspirante: Vinícius de Sousa Rodrigues

Primeiramente, fiz um entendimento geral da base de dados olhando os dados na planilha e anotei no primeiro tópico no Colab os tipos de variáveis (categóricas e numéricas). Daí dividi o trabalho em 3 etapas no tópico “Objetivos do desafio”: Passos base, Etapa de Padronização e Etapa de Enriquecimento.

- Etapa de Passos base

Nessa primeira etapa, eu importei a biblioteca pandas para manipular e analisar os dados e importei a base a ser trabalhada, armazenando-a na variável “base”. Depois visualizei as informações das colunas da base para saber como foi armazenado cada tipo de variável (float ou string). Por fim, verifiquei se havia dados nulos para que pudessem ser corrigidos e cheguei a conclusão de que não havia dados faltantes.

- Etapa de Padronização

Nessa segunda etapa, fiz os objetivos 1 e 2. Para o objetivo 1, criei e apliquei em cada linha da coluna “sexo” uma função para padronizar o sexo baseado na ideia que tive de que todo dado que se referia ao “Masculino” começava com a letra “m” ou “M” e que todo dado que se referia ao “Feminino” começava com a letra “f” ou “F”. Assim, a função verificava a letra inicial do dado e substituía pela sua correspondência, seja Masculino ou Feminino. Logo depois, printei alguns dados para verificar se as modificações tiveram sucesso.

Já para o objetivo 2, apenas substituí por meio da função replace de Python os pontos por vírgulas nos valores das colunas “nota_portugues” e “nota_matematica”, o que foi facilitado pelos dados dessas colunas estarem armazenados em string. Logo depois, printei alguns dados para verificar se as modificações tiveram sucesso.

- Etapa de Enriquecimento

Nessa terceira etapa, fiz os objetivos 3 e 4. Para o objetivo 3, criei uma variável temporária da base para que as modificações que eu fizesse não

prejudicassem a base original. Nessa base temporária, converti os valores das colunas “frequencia”, “nota_portugues” e “nota_matematica” para float para poder calcular a média com base na fórmula dada na documentação. Depois adicionei essa coluna “media” na base temporária com aproximação de duas casas decimais, printei alguns dados para verificar se as modificações tiveram sucesso e assim que vi que deu certo, transferi a coluna “media” da base temporária para a original.

Já para o objetivo 4, adicionei a coluna “aprovado” usando uma função de Python que verifica se o dado na linha correspondente da coluna “media” é maior/igual que 7, armazenando “Sim” na coluna “aprovado” dessa linha em caso positivo e “Não” em caso negativo. Logo depois, printei alguns dados para verificar se as modificações tiveram sucesso.

Por fim, fiz os mesmos passo-a-passo do objetivo 2 nas colunas “media” e “frequencia” e salvei a base atualizada, padronizada e enriquecida.

- Base atualizada:

https://drive.google.com/file/d/1sKQ-OifsTsq_sL8aJ6TxhTZ4XdwzANWA/view?usp=drive_link