descomplica
faculdade digital

Nome completo: Vinicius Dalla Zana de Mello
Curso: Pós-Graduação em Data Science
Área: **Nº do estudo de caso:**

Lembretes importantes:

- Leia o manual para elaboração de estudo de caso;
- Não é necessário reproduzir o enunciado do estudo de caso;
- Não se preocupe com a ABNT! Seu trabalho pode seguir este template (fonte Arial, tamanho 12 com espaçamento simples);
- O estudo de caso deve ter no mínimo 350 e máximo de 500 palavras contando a partir do título.

Análise de limpeza e tratamento de dados com Python

a. Para tratar os valores ausentes do DataFrame foi utilizado o método fillna(), que serve para preencher os valores ausentes em uma estrutura de dados. A equipe de análise de dados da ABC Company, utilizou o método mean() para preencher os valores da média na coluna “price” e o método median() para preencher os valores da mediana na coluna “sales”.
Para remover os valores ausentes, foi utilizado método dropna() do DataFrame pandas e o parâmetro “subset” para procurar por valores ausentes nas colunas ‘store_name’ e ‘date’, e assim remover as linhas. Agora, a variável df terá os valores atualizados.

b. Para corrigir erros de digitação e padronizar os nomes das lojas, foi criada a função clean_store_name com o parâmetro store_name. Na primeira linha da função temos a utilização do método strip(), que é usado para eliminar espaços em branco extras, que possam aparecer no inicio ou no final da string. Na segunda linha da função, foi utilizado o método replace(), que serviu para fazer a substituição de todas palavras Str. para Store, no nome da loja. Na terceira linha, temos o método replace() novamente, que substitui um espaço por outro, o que não faz sentindo nesse tratamento. Ao final, a função retorna o nome da loja com todas as alterações.
Para aplicar a função aos nomes das lojas, foi chamada a variável df[‘store_name’] recebendo-a com o método apply(), que serviu para aplicar a função a cada elemento da coluna store_name.

c. Para converter as datas e horas para um formato padrão, foi chamado a coluna date do df, recebendo a função do pandas pd.to_datetime(), que serviu para converter os valores da coluna date para objetos do tipo datetime. Também foi passado o parâmetro ‘erros’, definido como ‘coerce’, ou seja, caso a função encontrar algum valor na coluna ‘date’ que não possa ser convertido em ‘datetime’, ele ficará como ‘NaT’, que é a representação do pandas para valores de data e hora ausentes ou inválidos.

d. Para corrigir os preços negativos e valores de vendas, foi chamada a coluna ‘price’ e sales do df, recebendo-a com o método apply() e com a função anônima lambda, que neste caso, recebe um valor ‘x’ como entrada e retorna um valor absoluto, removendo o sinal negativo ou não alterando se for positivo.

1. Para instalar a biblioteca pandas, pode ser utilizado o gerenciador de pacotes pip, com o comando “pip install pandas” em um terminal ou no prompt.
   Para importar o pandas é utilizado o comando “import pandas as pd”.
   Para importar o dataframe pode ser utilizado o comando “df = pd.read_csv(‘dados.csv’)”.

2. Para remover valores duplicados, pode ser utilizado o comando “df = df.drop_duplicates(subset=[‘coluna’])”.

3. A função lambda em python, serve para criar pequenas funções anônimas em uma única linha de código, podendo ser usadas como argumentos para outras funções ou quando precisa ser criado uma função simples, que será usada apenas uma vez.