descomplica/pós

Estudo de Caso

Olá, estudante!

Se você está lendo este texto é porque você chegou à etapa de avaliação das disciplinas de Eixo do seu curso de Pós-graduação na área de Tecnologia.

Para desenvolver esta atividade, você deve ler os artigos selecionados abaixo e elaborar um texto com o limite de 500 palavras (utilize o contador de palavras do aplicativo Word ou equivalente para este controle). Não se esqueça de se posicionar com clareza a respeito dos aspectos técnicos e legais sobre os assuntos abordados. Ao final, um caso hipotético é apresentado para que você faça as suas considerações.

Artigos para leitura e fundamentação da análise

Estudo de Caso: Limpeza e Tratamento de Dados com Python

Introdução:

A ABC Company coleta e analisa dados de vendas de várias lojas em todo o país. Esses dados são usados para tomar decisões sobre estratégias de marketing e gerenciamento de estoque. No entanto, os dados coletados geralmente contêm informações incompletas, inconsistentes ou incorretas. Para melhorar a qualidade dos dados e obter insights mais precisos, a equipe de análise de dados da ABC Company decide realizar a limpeza e o tratamento dos dados utilizando Python.

A equipe de análise de dados da ABC Company começou analisando os dados brutos e identificou os seguintes problemas:

- Valores ausentes em algumas colunas.
- Dados duplicados.
- Erros de digitação nos nomes das lojas.
- Diferentes formatos de data e hora.
- Preços negativos e valores de vendas.

Logo definiram como objetivos da limpeza e tratamento dos dados o seguinte:

- Lidar com valores ausentes.
- Remover dados duplicados.
- Corrigir erros de digitação e padronizar os nomes das lojas.
- Converter datas e horas para um formato padrão.
- Corrigir preços negativos e valores de vendas.

descomplica/pós

Procedimento de Limpeza e Tratamento de Dados com Python:

A equipe decide usar a biblioteca pandas para a limpeza e o tratamento dos dados. Primeiro, instalaram a biblioteca pandas, importaram-na e importaram os dados do arquivo “dados.csv” para um *dataframe* do Pandas.

Depois utilizaram as etapas abaixo para limpar e tratar os dados:

a. Verificaram e trataram valores ausentes:

```
# Preencher valores ausentes com médias ou mediana (dependendo da distribuição)
df['price'] = df['price'].fillna(df['price'].mean())
df['sales'] = df['sales'].fillna(df['sales'].median())

# Remover linhas com valores ausentes em colunas específicas
df = df.dropna(subset=['store_name', 'date'])
```

b. Corrigiram erros de digitação e padronize os nomes das lojas:

```
# Função para corrigir erros de digitação comuns e padronizar os nomes das lojas
def clean_store_name(store_name):
    store_name = store_name.strip()
    store_name = store_name.replace("Str.", "Store")
    store_name = store_name.replace("  ", " ")
    return store_name

# Aplicar a função aos nomes das lojas
df['store_name'] = df['store_name'].apply(clean_store_name)
```

c. Converteram datas e horas para um formato padrão:

```
df['date'] = pd.to_datetime(df['date'], errors='coerce')
```

d. Corrigiram preços negativos e valores de vendas:

```
df['price'] = df['price'].apply(lambda x: abs(x))
df['sales'] = df['sales'].apply(lambda x: abs(x))
```

Problemática para desenvolvimento

(1) A equipe para poder tratar os dados precisou antes instalar a biblioteca Pandas, importar a biblioteca e importar o *dataframe*. Qual o código para instalar a biblioteca Pandas? E para importá-la? E para importar o dataframe?
(2) A equipe esqueceu de remover duplicados, como você pode fazer isso no Pandas?
(3) Em geral, para que serve a função *lambda* no python?