

Simulator Télos -

Nível #9 (Trilha de dados)

Seja muito bem-vindo ao Simulador - Nível #9! Neste nível, revisão alguns conceitos importantes sobre análise de dados utilizando diversas ferramentas e técnicas avançadas, como engenharia de dados, processos ETL e computação em nuvem. Aprendeu como lidar com grandes volumes de dados de forma eficiente e como transformar informações em insights valiosos. Está pronto para aplicar esse conhecimento em um projeto prático e criar um dashboard interativo que responda a questões estratégicas? 🚀

Contextualização

Imagine que você foi contratado para ser o responsável por analisar e interpretar as vendas de e-commerce de um influenciador renomado, que vende produtos para uma audiência diversa espalhada pelo mundo.

Seu objetivo é transformar dados brutos em insights valiosos, incluindo identificar tendências de vendas, compreender o perfil dos compradores e prever comportamentos futuros. Você deve apresentar essas descobertas de forma clara e visual em um **dashboard interativo**, que será utilizado pelo influenciador para tomar decisões estratégicas.

Será que você consegue mostrar o impacto dos envios internacionais, prever o desempenho de futuras vendas e revelar os produtos mais populares? Bem-vindo ao desafio do Simulador - Nível #9!

Descrição do Simulador

Os **Dados Globais de Vendas no Varejo** fornecidos são um dataset sintético gerado por técnicas de amostragem aleatória disponibilizadas pelo pacote Numpy. O dataset simula informações sobre vendas de mercadorias realizadas por meio de um site de varejo gerido por um influenciador fictício popular, baseado nos Estados Unidos, no período entre 2023 e 2024.

O influenciador vende roupas, ornamentos e outros produtos a preços variáveis por meio do site para seguidores em todo o mundo. Imagine que o influenciador realiza altos níveis de promoções para os produtos vendidos, incentivando mais avaliações e resenhas de seus seguidores, o que aumenta o engajamento dos usuários.

Descrição das colunas:

- **Order ID:** Identificador único de cada pedido.
- **Order Date:** Data em que o pedido foi feito.
- **Product ID:** Identificador único do produto vendido.
- **Product Category:** Categoria do produto vendido (Roupas, Ornamentos, Outros).
- **Buyer Gender:** Gênero dos compradores (Masculino, Feminino).
- **Buyer Age:** Idade dos compradores.
- **Order Location:** Cidade onde o pedido foi realizado.
- **International Shipping:** Indica se o produto foi enviado internacionalmente (Sim/Não).
- **Sales Price:** Preço do produto.
- **Shipping Charges:** Taxas extras para envios internacionais.
- **Sales per Unit:** Custo de venda incluindo as taxas de envio internacional.
- **Quantity:** Quantidade de produto comprada.
- **Total Sales:** Total de vendas realizadas na compra.
- **Rating:** Avaliação do usuário para o pedido.
- **Review:** Resenha dada pelo usuário para o pedido.

Você deverá criar um **dashboard interativo** que responda a questões estratégicas relacionadas ao dataset e ofereça insights visuais claros sobre o desempenho das vendas e as preferências dos clientes. Para chegar a esse resultado, será preciso desenvolver o ELT.

Histórias de Usuário a serem Implementadas

1) ETL

1. Extração

- **Importar os dados:** O primeiro passo é importar o dataset necessário para o Google Colab.

```
import kagglehub
import pandas as pd
import os
# Download
path = kagglehub.dataset_download("adarsh0806/influencer-merchandise-sales")

# Listar todos os arquivos no diretório do conjunto de dados
files = os.listdir(path)
print("Arquivos no dataset:", files)

# Lendo arquivos CSV no conjunto de dados
for file in files:
    if file.endswith(".csv"):
        csv_path = os.path.join(path, file)
        # Carregar o arquivo CSV em um DataFrame
        data = pd.read_csv(csv_path)
        print("Dataset carregado com sucesso!")
        print(data.head())
        break
```

```
else:  
    print("Nenhum arquivo CSV foi encontrado.")
```

2. Transformação

- Limpeza dos dados: remover ou corrigir dados ausentes; converta colunas de dados de texto para numéricas, se necessário.
- Criação de novas variáveis: calcular novas métricas, como lucro por venda.
- Incluir dados do modelo de previsão

3. Carregamento

- Enviar os dados para o Google Spreadsheet.

2) Conecte o Looker studio a planilha do Google gerada para criar o dashboard.

O dashboard deverá incluir:

1. Análise de Vendas e Tendências

- Quais categorias de produtos geraram maior receita ao longo do tempo?
- Há sazonalidade nas vendas ao longo do período analisado?

2. Comportamento dos Compradores

- Qual é o perfil médio dos compradores (idade e gênero)?
- Como as avaliações e resenhas variam entre os produtos? **(opcional: desafio 1)**

3. Desempenho Geográfico

- Quais cidades ou regiões geraram o maior volume de vendas?
- Os pedidos internacionais contribuíram significativamente para as vendas totais?

4. Previsão e Análise de Decisão (Regressão Logística)

- Prever se as vendas para o mês seguinte serão superiores ou inferiores com relação ao período presente, utilizando Regressão Logística.

3) Cloud computing (Opcional)

Utilize o Google Bigquery em vez do Colab. Carregue o csv construindo uma tabela, e utilize SQL para processar os dados. Crie uma tabela final com os resultados processados. Realize a conexão do Looker diretamente nessa tabela do Bigquery.

Critérios de Aceitação:

- ☰ Entrega do código usado no ETL (Git, ou colab)
- ☰ Base gerada na planilha do Google;
- ☒ Dashboard com as métricas apresentadas em elementos gráficos e/ou tabelas.

Importante manter os recursos gerados com acesso público.

- Link para ajudar no desenvolvimento gráfico no Looker:
<https://support.google.com/looker-studio/answer/12268451?hl=pt-BR>
- Link para a utilização do Google Bigquery sem custo:
<https://cloud.google.com/bigquery/docs/sandbox?hl=pt-br>

Cronograma de Desenvolvimento Proposto

- ☒ **Dia 1 e 2:** Desenvolvimento da História de Usuário 1.
- ☒ **Dia 3:** Desenvolvimento da História de Usuário 2.
- ☒ **Dia 4:** História de Usuário 4 (opcional).

Avaliação

Conteúdo e Precisão:

- O dashboard apresenta informações relevantes e bem relacionadas ao tema do dataset?
- Os dados são precisos e manipulados corretamente?

Análise e Insight:

- Realizou uma análise adequada das tendências e comportamentos dos dados?
- As relações entre variáveis são claras e evidentes?

Visualização e Apresentação:

- O dashboard é visualmente atraente e fácil de entender?
- Gráficos são adequados para as análises realizadas?

Qualidade do Modelo Preditivo (se aplicável):

- O modelo preditivo é adequado para os dados e atinge os objetivos propostos?

Competências Avaliadas

Entendimento de Análise Exploratória de Dados (EDA):

- Capacidade de realizar análise exploratória para entender a estrutura e as relações nos dados.
- Conhecimento sobre como identificar padrões e tendências nos dados iniciais.

Pré-processamento de Dados:

- Familiaridade com técnicas de pré-processamento como normalização, tratamento de valores ausentes e codificação de variáveis.

Conhecimento em Algoritmos de Machine Learning Supervisionado:

- Entendimento de algoritmos como regressão linear.

Habilidade em Programação com Python:

- Capacidade em manipular dados com pandas.

Organização do Código e Boas Práticas:

- Capacidade de organizar o código de forma clara e reproduzível, utilizando boas práticas.

Integração de ferramentas