Projeto de Avaliação - Laboratório de Ciências de Dados

Objetivo

Este projeto tem como objetivo avaliar as habilidades dos alunos em manipulação, limpeza e análise de dados utilizando um banco de dados real. O dataset escolhido para este projeto é o **Grupo Bimbo Inventory Demand**, que oferece um desafio complexo e realista, semelhante ao que os alunos podem encontrar em um ambiente de trabalho.

Neste projeto, não serão fornecidas 'questões' específicas a serem respondidas. Por exemplo, não haverá perguntas do tipo "Qual é a média de vendas por semana?" ou "Remova os valores ausentes da coluna x". Em vez disso, os alunos devem explorar o dataset, levantar hipóteses, realizar análises e criar visualizações para comunicar os resultados.

Descrição do Dataset

Bimbo é uma empresa mexicana de panificação que atua em diversos países da América Latina. O dataset do desafio **Grupo Bimbo Inventory Demand** contém informações detalhadas sobre as vendas e devoluções de produtos da Bimbo em diversas regiões e rotas de entrega no período de 9 semanas.

- `Semana`: Semana do registro da venda.
- `Agencia_ID`: Identificador do depósito da Bimbo.
- `Canal_ID`: Canal de venda.
- `Ruta_SAK`: Identificador da rota.
- `Cliente_ID`: Identificador do cliente.
- `Producto_ID`: Identificador do produto.
- `Venta_uni_hoy`: Unidades vendidas na semana.
- `Venta_hoy`: Valor monetário das vendas na semana.
- `Dev_uni_proxima`: Unidades devolvidas na próxima semana.
- `Dev_proxima`: Valor monetário das devoluções na próxima semana.
- `Demanda_uni_equil`: Demanda ajustada do produto.

> `Demanda_uni_equil` é a principal variável de interesse neste *dataset*. Ela representa a demanda ajustada do produto, que é a quantidade real de produtos que os clientes compraram. Isso porque o prejuízo causado por excesso ou falta de produtos nas prateleiras é do interesse da Bimbo.

Etapas do Projeto

1. Exploração Inicial e Entendimento do Dataset

- **Tarefa**: Os alunos devem realizar uma análise inicial para entender a estrutura e as características do dataset.
- **Objetivo**: Compreender as diferentes colunas e o que cada uma representa no contexto do negócio.

> **Dica**: Foi disponibilizado um arquivo `train_sample.csv` com uma amostra do dataset completo. Utilize-o para a exploração inicial e, posteriormente, aplique as mesmas técnicas ao dataset completo.

> **Dica**: Não é necessário descompactar o arquivo `train.csv.zip`. Você pode carregar o arquivo diretamente com a biblioteca `pandas`.

```
>>> import pandas as pd
>>> df = pd.read_csv('train_sample.csv.zip')
```

Junto ao *dataset* completo, foi disponibilizado um arquivo `cliente_tabla.csv` com informações sobre os clientes, e um arquivo `producto_tabla.csv` com informações sobre os produtos. Esses arquivos podem ser utilizados para enriquecer a análise.

2. Limpeza de Dados

- **Tarefa**: Realizar a limpeza dos dados, que inclui o tratamento de valores ausentes, duplicatas e inconsistências.
- **Objetivo**: Garantir que os dados estejam em um estado adequado para análise, minimizando erros e vieses.

3. Análise Exploratória de Dados (EDA)

- **Tarefa**: Realizar uma análise descritiva e exploratória dos dados.
- **Objetivo**: Identificar padrões, tendências e possíveis relações entre variáveis, como vendas, devoluções, regiões e produtos.

> **Dica**: Elabore as perguntas como se não fosse você mesmo a respondê-las. Primeiro, elabore as perguntas e, em seguida, busque formas de respondê-las, não ao contrário, i.e., não pense "o que dá para saber agrupando esta variável por aquela?", mas sim "o que eu gostaria de saber sobre esta variável?".

4. Visualizações

- **Tarefa**: Criar visualizações que destaquem os principais *insights* descobertos na análise exploratória.
- **Objetivo**: Utilizar gráficos de séries temporais, mapas de calor, e gráficos de dispersão para comunicar de forma clara as descobertas.

> Dica: Essa etapa não precisa ser feita de forma linear. Você pode criar visualizações ao longo da análise exploratória e refazê-las conforme novas perguntas surgirem. As etapas 3 e 4 pode ser feitas de forma iterativa.

5. Relatório e Apresentação

- **Tarefa**: Documentar todo o processo, desde a exploração inicial até as conclusões finais, no formato de um relatório *jupyter notebook* ou *colab notebook*. Cada trecho do código deve ser acompanhado de uma explicação clara e objetiva, constando as perguntas que motivaram a análise e as respostas encontradas. A parte textual deve ser escrita com a formatação adequada, incluindo títulos, subtítulos, parágrafos e listas em *markdown*.
- **Identificação**: O relatório deve conter nome e matrícula dos alunos, e o nome do projeto.
- **Objetivo**: Desenvolver a capacidade de comunicação dos alunos, essencial para a atuação em ciência de dados.

Critérios de Avaliação

- **Qualidade da Limpeza de Dados** (25%): Avaliar como os alunos lidaram com os dados brutos e transformaram-nos em um conjunto de dados utilizável.
- **Profundidade da Análise Exploratória** (35%): Avaliar a habilidade dos alunos em identificar e interpretar padrões significativos nos dados.
- **Qualidade das Visualizações** (20%): Avaliar a clareza, eficácia e **beleza** das visualizações criadas.
- **Documentação e Apresentação** (20%): Avaliar a organização e clareza da documentação e da apresentação final.

> **Nota**: A beleza das visualizações é um critério importante. Visualizações claras e atraentes são essenciais para comunicar eficazmente os resultados da análise.

Entrega

- **Formato**: O relatório final deve ser entregue *online* no formato de um *notebook* (Jupyter ou Colab).
- **Equipes**: O projeto deve ser feito em equipes de até 5 alunos.
- **Prazo**: dia 19/09/2024 às 23:59.