Data Summary Report - RetailSense AI

Projeto: RetailSense AI

Descrição:

Este relatório resume as fontes de dados e as principais variáveis que serão utilizadas no projeto RetailSense AI, com o objetivo de analisar dados de transações de varejo e gerar insights sobre oportunidades de negócio e práticas sustentáveis.

Fontes de Dados Utilizadas:

1. **Online Retail Dataset** (UCI Machine Learning Repository):
    - Fonte: https://archive.ics.uci.edu/dataset/352/online+retail
    - Descrição: Dados transacionais de um varejista online do Reino Unido, contendo todas as transações ocorridas entre 01/12/2010 e 09/12/2011. A empresa vende principalmente presentes únicos para todas as ocasiões, e muitos de seus clientes são atacadistas.

Variáveis Originais:

1. **InvoiceNo**: Número único da fatura por transação.
2. **StockCode**: Código do produto.
3. **Description**: Descrição do produto.
4. **Quantity**: A quantidade de cada produto por transação.
5. **InvoiceDate**: Data e hora da fatura.
6. **UnitPrice**: Preço unitário do produto em libras esterlinas.
7. **CustomerID**: ID único do cliente.
8. **Country**: Nome do país.

Processamento dos Dados e Feature Engineering:

- Os dados originais estavam em formato XLSX, mas foram convertidos para CSV para facilitar o manuseio.
- Os dados foram processados, tratados e limpos antes de serem armazenados em um banco de dados Supabase.

- Foram realizadas transformações e criadas novas variáveis por meio de feature engineering para enriquecer a análise, resultando nas seguintes variáveis finais:

Variáveis Finais (Após Tratamento e Feature Engineering):

1. **NumeroFatura**: Número único da fatura por transação (renomeado de InvoiceNo).
2. **CodigoProduto**: Código do produto (renomeado de StockCode).
3. **Descricao**: Descrição do produto (renomeado de Description).
4. **Quantidade**: A quantidade de cada produto por transação (renomeado de Quantity).
5. **DataFatura**: Dia, mês e ano da fatura (extraído de InvoiceDate).
6. **PrecoUnitario**: Preço unitário do produto em libras esterlinas (renomeado de UnitPrice).
7. **IDCliente**: ID único do cliente (renomeado de CustomerID).
8. **Pais**: Nome do país (renomeado de Country).
9. **CategoriaProduto**: Categoria do produto (criada com base na Descricao).
10. **CategoriaPreco**: Categoria de preço - Barato, Moderado, Caro (criada com base no PrecoUnitario).
11. **Ano**: Ano da transação (extraído de DataFatura).
12. **Mes**: Mês da transação (extraído de DataFatura).
13. **Dia**: Dia da transação (extraído de DataFatura).
14. **DiaSemana**: Dia da semana da transação - 0-6, sendo 0 = Segunda-feira (extraído de DataFatura).
15. **SemanaAno**: Semana do ano da transação - 1-52 (extraído de DataFatura).
16. **ValorTotalFatura**: Valor total da fatura (calculado como Quantidade * PrecoUnitario).
17. **FaturaUnica**: Se a fatura é única (True) ou faz parte de uma fatura maior (False) (criada com base em NumeroFatura).

- Os dados limpos e enriquecidos estão armazenados em duas tabelas no banco de dados Supabase:
    - **transactions_main**: Contém todos os dados processados (mais de 500 mil linhas).
    - **transactions_sample**: Contém uma amostra aleatória e representativa de 50 mil linhas da tabela "transactions_main". Esta tabela será utilizada para análise e prototipação no projeto RetailSense AI, devido a questões de desempenho ao trabalhar com o conjunto de dados completo no Streamlit.

Observações:

- **Dados Abertos**: O conjunto de dados utilizado é aberto e permitido para uso e análise por qualquer pessoa.
- **Amostragem de Dados**: Para fins de prototipação e desempenho, será utilizada a tabela "transactions_sample" com 50 mil linhas selecionadas aleatoriamente da tabela "transactions_main". Isso permitirá uma análise mais eficiente no Streamlit.
- **Evolução do Projeto**: Novas fontes de dados, especialmente dados de sustentabilidade, poderão ser integradas conforme o projeto evolui para aprimorar ainda mais os insights gerados.