**RETAIL SALES DATASET**

O presente trabalho terá a seguinte estrutura:

1.Objetivo

2. Dataset

3. Coleta

4. ETL

5. BigQuery

Créditos: O notebook foi desenvolvido e documentado por:

* Cleiton Silva

**1. OBJETIVO**

A partir dos dados compreender como a idade e o sexo dos clientes influenciam seus padrões de compra, Identificar quais categorias de produtos são mais populares entre os clientes, Investigar as relações entre a idade dos clientes, seus gastos e as categorias de produtos preferidas.

**2. DATASET**

Os dados Retail Sales Dataset foram retirados do site do Kaggle, Este conjunto de dados representa uma simulação de um ambiente de varejo imaginário, abrangendo elementos críticos que afetam as operações de varejo e os relacionamentos com os clientes. Ele engloba informações essenciais, como identificação de transações, datas, identificação de clientes, gênero, faixa etária, categoria de produtos, quantidade, preço unitário e valor total. Esses parâmetros possibilitam uma análise abrangente das tendências de vendas, influências demográficas e padrões de compra.

Esses dados foram retirados do Kaggle através do seguinte link: <https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset>.

O dataset é dividido da seguinte forma:

* Transaction ID - Um identificador exclusivo para cada transação, permitindo rastreamento e referência.
* Date - A data em que a transação ocorreu, fornecendo insights sobre as tendências de vendas ao longo do tempo.
* Customer ID - Um identificador exclusivo para cada cliente, permitindo uma análise centrada no cliente.
* Gender - O gênero do cliente (masculino/feminino), oferecendo insights sobre padrões de compra baseados em gênero.
* Age - A idade do cliente, facilitando a segmentação e exploração de influências relacionadas à idade.
* Product Category - A categoria do produto adquirido (por exemplo, Eletrônicos, Roupas, Beleza), ajudando a entender as preferências do produto.
* Quantity - Quantidade de unidades do produto adquiridas, contribuindo para insights sobre volumes de compras.
* Price per Unit - O preço de uma unidade do produto, auxiliando nos cálculos relativos ao gasto total.
* Total Amount - O valor monetário total da transação, mostrando o impacto financeiro de cada compra.

2.1 FLUXOGRAMA DO PROJETO

O projeto terá seu início com a retirada dos dados do Kaggle e inserido na nuvem do Google Cloud Platform, o Data Fusion ficará responsável pelo ETL e será finalizado no BigQuery. Abaixo imagem exemplificando.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

**3. COLETA**

Dados baixados para a máquina local e inseridos manualmente no bucket do Google Cloud Storage.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**4. ETL**

Os dados foram retirados do bucket e feitas as seguintes alterações:

1º Split da coluna data criando 3 novas colunas dia, mês e ano; e

2º Removi a coluna Customer\_ID não vi finalidade para o atual projeto.

Diagrama

Descrição gerada automaticamente

**4. BIGQUERY**

Através das queries, respondi as perguntas do início.

* Como podemos comparar o comportamento de compra entre diferentes grupos etários e gêneros?

**Query**

Texto

Descrição gerada automaticamente

**Resultado**

Tabela

Descrição gerada automaticamente

Com base no resultado apresentado, tive à seguinte conclusão?

No perfil de 18 a 20 anos tanto as meninas quanto os meninos utilizam muito eletrônicos;

No perfil de 21 a 30 anos as meninas passam a consumir mais produtos de beleza, enquanto os meninos preferem gastar com roupas;

No perfil de 31 a 40 anos os gostos passam a se igualar novamente, isso porque os homens mantem o seu consumo em roupas e as mulheres passam a consumir a mesma categoria.

No perfil de 41+ os gostos voltam a alterar, porque as mulheres mantem o consumo por roupas enquanto os homens voltam para os eletrônicos.

* Qual é o ranking das categorias com base na quantidade vendida ou na receita gerada?

**Query**

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Resultado**

Tabela

Descrição gerada automaticamente

O consumo de roupa foi o grande campeão, isso porque temos uma concentração de consumo dos homens no faixa etária de 21 a 40 anos e as mulheres acima de 30 anos.

* Qual a correlações entre idade, gastos e categorias de produtos compradas.

**Query**

Texto

Descrição gerada automaticamente com confiança baixa

**Resultado**

**Tabela

Descrição gerada automaticamente**

Tirando a coluna Cor\_Spendin\_Price não localizei outra com uma correlação alta.