**PLANEJAMENTO MVP – E-Commerce**

ETAPAS: Definir objetivo, definir perguntas, realizar a busca e coleta de dados, modelar os dados, carregar os dados e por fim, analisar os dados.

OBJETIVO: Desenvolver um pipeline de dados na nuvem para monitorar, analisar e prever o comportamento dos consumidores em uma plataforma de e-commerce. O objetivo é criar um MVP que ajude a melhorar a experiência do usuário e aumentar as vendas.

PERGUNTAS:

- Quais produtos estão tendo maior demanda em diferentes períodos?
- Quais são os padrões de comportamento de compra dos clientes?
- Quais fatores influenciam a decisão de compra dos consumidores?
- Quais produtos são frequentemente comprados juntos?
- Quais são as tendências de vendas futuras com base nos dados históricos?
- Qual categoria é mais comprada para cada gênero?

COLETA DE DADOS: Conjuntos de dados do Kaggle, como "E-Commerce Data".

Catálogo de dados: **Customer ID** (identificador único para cada cliente), **Purchase Date** (data em que foi feita cada compra de cada cliente), **Product Category** (categoria do produto comprado), **Product Price** (preço do produto comprado), **Quantity** (quantidade de produto comprado), **Total Purchase Amount** (total gasto por cliente em cada transação), **Payment Method** (método de pagamento), **Customer Age** (idade do cliente), **Returns** (se foi retornado o produto ou não (0 para não, 1 para sim)), **Customer Name** (nome do cliente), **Age** (idade do cliente), **Gender** (gênero do cliente), **Churn** (se houve ou não churn (1 para sim, 0 para não)).

MODELAGEM:

Irei modelar com o intuito de retificar modelos logicamente incorretos e minimizar redundâncias desses modelos.

TABELA INICIAL:

| Customer ID | Purchase Date | Product Category | Product Price | Quantity | Total Purchase Amount | Payment Method | Customer Age | Returns | Customer Name | Age | Gender | Churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 44605 | 03/05/2023 21:30 | Home | 177 | 1 | 2427 | PayPal | 31 | 1.0 | John Rivera | 31 | Female | 0 |
| 44605 | 16/05/2021 13:57 | Electronics | 174 | 3 | 2448 | PayPal | 31 | 1.0 | John Rivera | 31 | Female | 0 |
| 44605 | 13/07/2020 06:16 | Books | 413 | 1 | 2345 | Credit Card | 31 | 1.0 | John Rivera | 31 | Female | 0 |
| 44605 | 17/01/2023 13:14 | Electronics | 396 | 3 | 937 | Cash | 31 | 0.0 | John Rivera | 31 | Female | 0 |
| 44605 | 01/05/2021 11:29 | Books | 259 | 4 | 2598 | PayPal | 31 | 1.0 | John Rivera | 31 | Female | 0 |
| 13738 | 25/08/2022 06:48 | Home | 191 | 3 | 3722 | Credit Card | 27 | 1.0 | Lauren Johnson | 27 | Female | 0 |

1a Forma Normal: Deixar valores atômicos.

- Purchase Date será dividida em coluna data e coluna hora. Excluirei a coluna hora para fins de eficiência do modelo.
- Excluirei a coluna age que está duplicando a coluna “customer age”.
- Não removi linha duplicada, pois não havia nenhuma.
- Há de ter uma dependência funcional de todos os atributos em relação à chave primária.
- Não pode conter valores nulos em nenhuma das colunas e linhas.

Ao fim de todas as formas normais:

**Tabela Fato**:

Customer ID (índice, integer)

Purchase Date (date)

Product Category ID (índice, integer, a partir de 1)

Quantity (integer)

Total Purchase Amount (float)

**Time Dimension**:

Purchase Date

Day (integer)

Month (integer)

Year (integer)

**Customer Dimension**:

Customer ID

Customer Age (integer)

Customer Name (string)

Gender (string)

Payment Method (string)

Churn (integer, 0 ou 1)

Returns (integer, 0 ou 1)

**Category Dimension:**

Product Category ID (Index)

Product Category (string)

Product Price (float)

TRATAMENTO DOS DADOS:

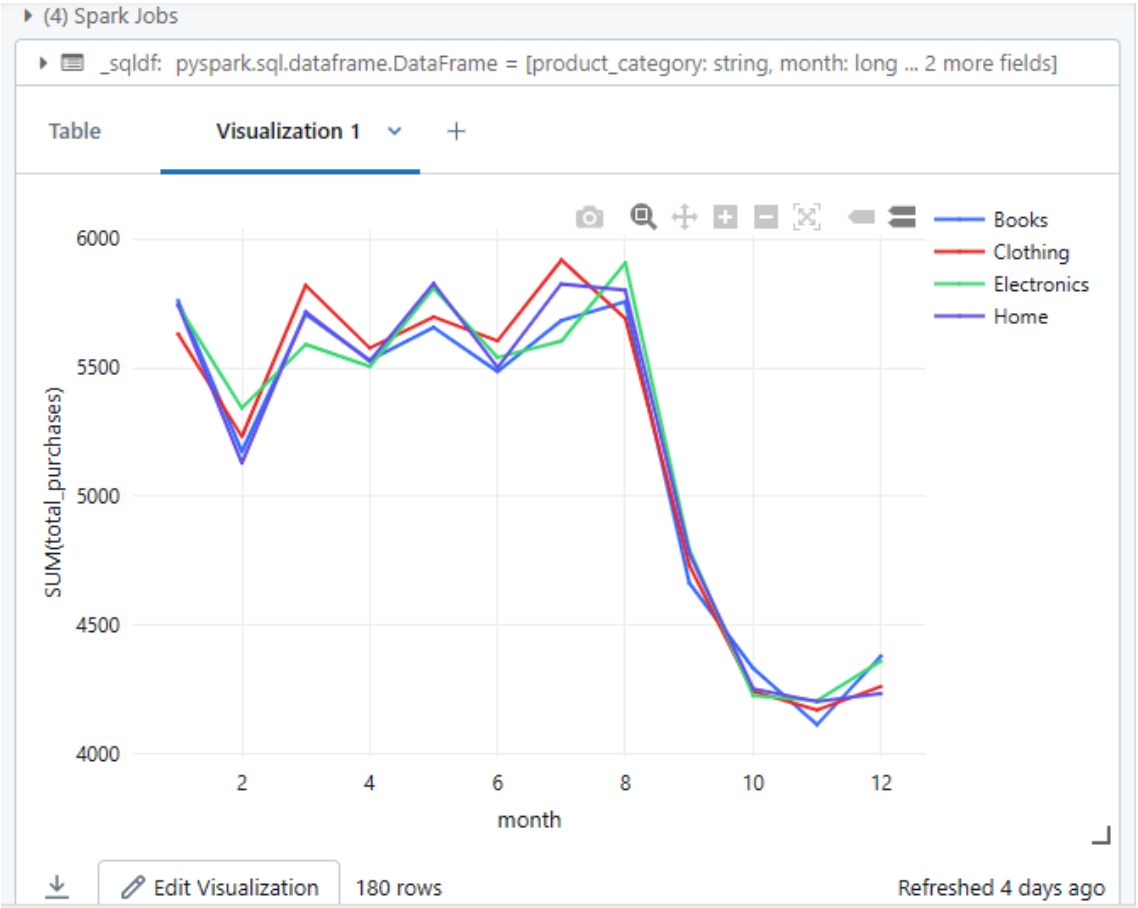
Para o tratamento dos dados, optei por realizar essa etapa no colab com os seguintes intuitos: Regularizar coluna data, retirar coluna “Age” duplicada, substituir nulos por 0 e transformar nomes compostos de colunas por nomes simples. Todas as transformações estão documentadas no documento “Tratamento dos dados MVP”. Por fim, criei as tabelas de dimensões e fato. Subi elas no github e copiei o “path” para o databricks para análise dos dados.

ANÁLISE DOS DADOS

As análises foram feitas no databricks com auxílio de queries e visualizações da plataforma, em gráficos.

**P1: Quais produtos estão tendo maior demanda em diferentes períodos?**

Para esta pergunta, obtive como resultado da query o seguinte gráfico:



Podemos ver que os resultados estão bastante alinhados entre si, com crescimento e caídas semelhantes. Analisando a tabela, é possível concluir que eletrônicos e roupas são os mais requisitados durante os meses de 2020. Esta análise nos auxilia a interpretar a quinta pergunta também: **P5: Quais são as tendências de vendas futuras com base nos dados históricos?** Podemos ver que a tendência, visto o comportamento de vendas dos últimos 4 anos, é as vendas aumentarem no início do ano e seguir variando até o mês de agosto. Isto pode ser resultado de um efeito sazonal, pela interpretação do gráfico, ou mesmo da má qualidade da base de dados.

**P2: Quais são os padrões de comportamento de compra dos clientes?**

Para a segunda pergunta realizei uma query que avalia as compras dos clientes que não retornaram nem desistiram do produto. Para cada cliente, pedi uma média da compra, o valor da compra máxima e mínima, assim como a soma de todas as compras feitas pelo cliente. Com esses dados, consigo observar o padrão de compra de cada cliente em específico, facilitando o processo de venda futura.

Complementando a P2, decidi observar qual produto é mais devolvido pelos clientes. Com isso, realizando um query, é possível realizar a análise de cada produto devolvido por cada cliente.

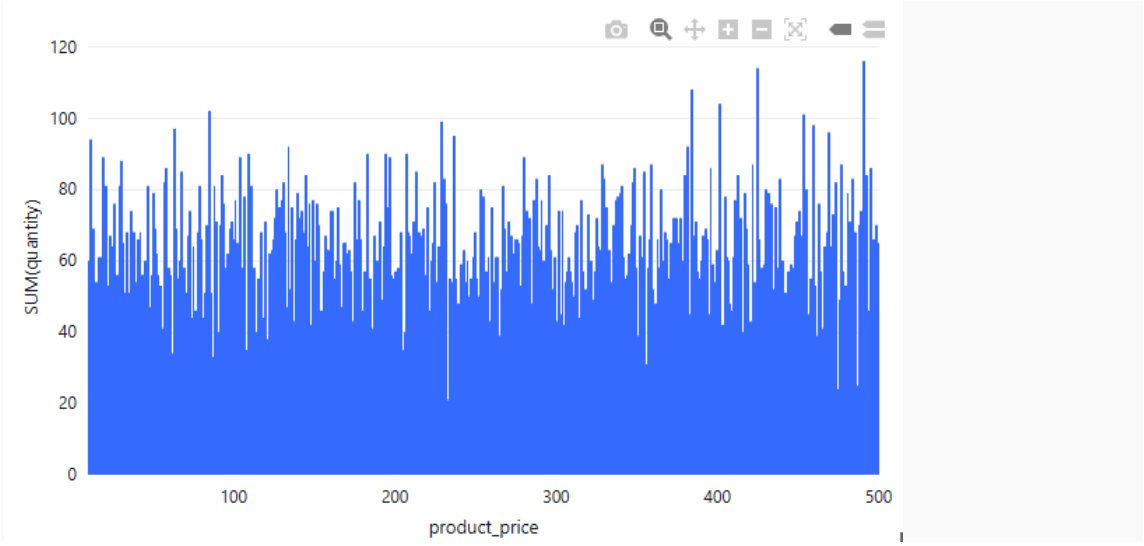
Destrinchando ainda mais a pergunta, decidi analisar quem é o cliente que mais devolve produtos. Chegando à conclusão de que é o Michael Smith.

Por fim, fiz uma query para descobrir qual o produto mais devolvido ou desistido. Sendo esse o da classe de eletrônicos.

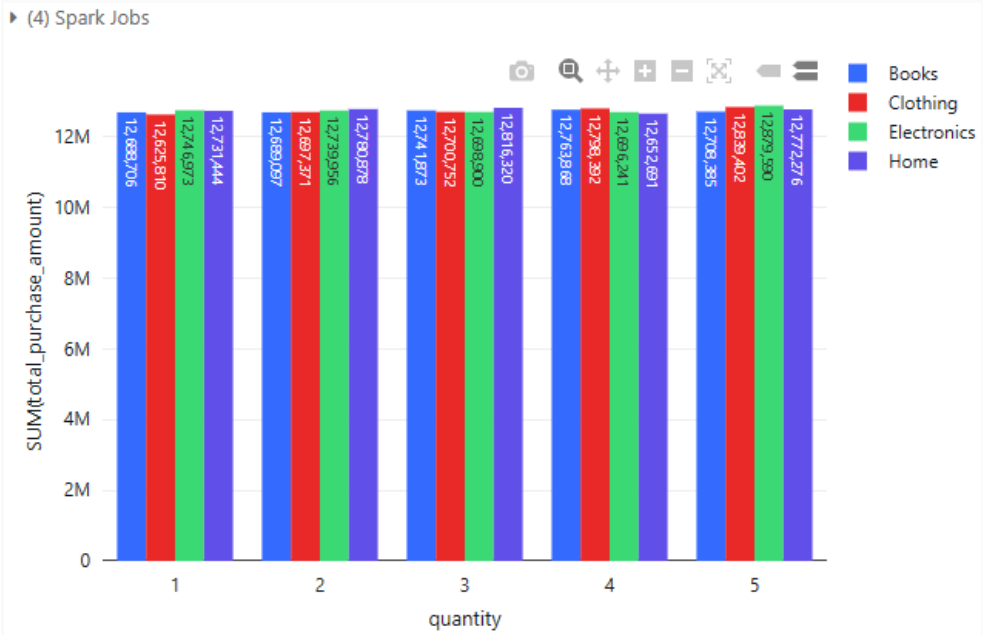
**P3: Quais fatores influenciam a decisão de compra dos consumidores?**

Para esta pergunta avaliei alguns fatores:

- `PASSO 1: Preço do Produto vs. Total de Compras: Analisa a relação entre o preço médio dos produtos e o valor total gasto, agrupado pela categoria do produto.`



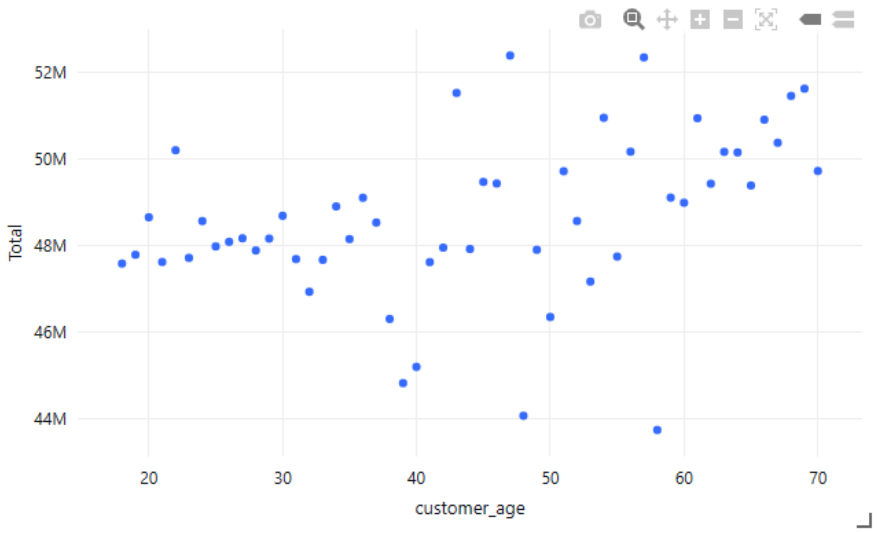

- PASSO 2: Categoria do Produto vs. Total de Compras: Examina quais categorias de produtos têm maior demanda, medido pelo valor total gasto. PRIMEIRO VOU CHECAR SE A CATEGORIA E A QUANTIDADE POSSUEM RELAÇÃO.




- PASSO 2: Categoria do Produto vs. Total de Compras: Examina quais categorias de produtos têm maior demanda, medido pelo valor total gasto. POR FIM VOU CHECAR SE A CATEGORIA E O VALOR TOTAL DA COMPRA POSSUEM RELAÇÃO.

| | Total | product_category |
|---|---|---|
| 1 | 171138916 | Home |
| 2 | 170716122 | Clothing |
| 3 | 170146025 | Electronics |
| 4 | 169345236 | Books |

- PASSO 3: Idade do Cliente vs. Total de Compras: Verifica como a idade do cliente está relacionada ao total gasto.




```
-- COMPLEMENTANDO A ANÁLISE DA PERGUNTA ANTERIOR, DECIDI CHECAR A MÉDIA DAS COMPRAS
```

- PASSO 4: Gênero do Cliente vs. Total de Compras: Analisa a influência do gênero do cliente no valor total gasto.

| | Gender | Total |
|---|---|---|
| 1 | Male | 1294942936 |
| 2 | Female | 1282800824 |

- PASSO 5: Forma de Pagamento vs. Total de Compras: Analisa a influência do método de pagamento do cliente no valor total gasto.

| | Pagamento | Total |
|---|---|---|
| 1 | Credit Card | 861209602 |
| 2 | PayPal | 859849684 |
| 3 | Cash | 856684474 |

É possível concluir que a idade é de fato um fator, pois clientes mais velhos possuem maior gastos registrados. Por uma pequena diferença, concluo que o método de pagamento também é um fator, visto à procura pelo credit card. Já as demais hipóteses não possuem resultados claros.

**P4: Quais produtos são frequentemente comprados juntos?**

Para essa análise, realizei uma query com o seguinte resultado:

| | product1_name | product2_name | count |
|---|---|---|---|
| 1 | Clothing | Clothing | 1 |
| 2 | Clothing | Home | 1 |
| 3 | Clothing | Books | 1 |
| 4 | Books | Electronics | 1 |
| 5 | Clothing | Clothing | 1 |
| 6 | Books | Electronics | 1 |
| 7 | Electronics | Books | 1 |
| 8 | Electronics | Books | 1 |
| 9 | Books | Books | 1 |
| 10 | Books | Clothing | 1 |

**P6: Qual categoria é mais comprada para cada gênero?**

Conclui-se que homens preferem itens de casa, assim como as mulheres.

| | Gender | product_category | Total |
|---|---|---|---|
| 1 | Male | Home | 324469092 |
| 2 | Male | Clothing | 324235310 |
| 3 | Male | Electronics | 323260216 |
| 4 | Male | Books | 322978318 |
| 5 | Female | Home | 322707660 |
| 6 | Female | Electronics | 321295140 |
| 7 | Female | Clothing | 320783341 |
| 8 | Female | Books | 318014683 |

**CONCLUSÃO**

Com este trabalho, conclui-se que, dentro da base de dados analisada, há diversos fatores que podem auxiliar a garantir a melhoria da experiência do usuário e aumento das vendas. Analisou-se diversos pontos, porém a base parece ser um tanto quanto montada, pois os valores estavam muito conformes, parecidos, com poucas divergências entre si. Isso impediu que muitos fatores viessem a se tornar um influenciador nas vendas, porém é importante ter em mente que é uma base teste e que as mesmas queries devem ser realizadas para uma nova base de dados. Desta forma, poderão aparecer novos influenciadores, fatores.

Deve-se, portanto, seguir o passo a passo da pipeline realizada para uma nova base de dados. Assim, garantir-se-á um bom planejamento para aumento de vendas.

Para fins de aprendizado, posso afirmar que o trabalho em questão me auxiliou muito na descoberta e conhecimento da plataforma databricks, assim como no desenvolvimento de minhas skills de programação e análise de dados. Embora não tenha sido criado um dash atrelado à base de dados, consegui realizar diversas análises por meio das visualizações do próprio databricks.

Minhas maiores dificuldades rondaram as questões práticas, isto é, como manusear a plataforma. Dificuldade, esta, combatida com maestria por meio do auxílio acadêmico da faculdade, permitindo que eu concluísse o projeto final. Para trabalhos futuros na nova sprint, espero conhecer um pouco mais sobre as visualizações e o poder de análise que elas têm a fornecer.