

Visão Geral

Este projeto apresenta, de forma completa, todas as principais etapas do trabalho de um analista de dados — desde a preparação e limpeza do dataset, passando pela criação de consultas SQL para exploração aprofundada, até a construção de dashboards interativos no Looker Studio.

A análise foi realizada sobre um conjunto de mais de **3.900 compras** distribuídas em diversas categorias de produtos, permitindo uma visão ampla do comportamento de compra dos clientes. O objetivo central foi identificar **padrões de gastos**, segmentar clientes com base em seu histórico, compreender **preferências por tipos de produtos** e analisar **tendências de assinatura e recorrência**.

Ao longo do estudo, buscou-se transformar dados brutos em **insights açãoáveis**, revelando oportunidades de otimização, estratégias de retenção e compreensão mais clara do perfil de consumo dos diferentes segmentos de clientes. O resultado é um projeto robusto, visualmente organizado e guiado pela metodologia analítica do início ao fim.

Sobre o Conjunto de Dados

O conjunto de dados utilizado neste projeto possui **3.900 linhas e 18 colunas**, reunindo informações completas sobre clientes, compras e padrões de consumo. Ele inclui campos como **ID do cliente, idade, gênero, item comprado, categoria do produto, valor da compra, localização, tamanho, cor, estação, avaliação do cliente, status de assinatura, tipo de envio, desconto aplicado, compras anteriores, método de pagamento, frequência de compras, faixa etária e frequência de compra em dias**.

Essa variedade de atributos permite uma análise profunda do comportamento dos consumidores, abrangendo desde características demográficas até preferências de produtos, padrões de recorrência, impacto de descontos e tendências de assinatura. É um dataset rico, ideal para gerar insights e embasar decisões estratégicas ao longo do projeto.

Análise Exploratória com Python

Comecei os preparativos limpando os dados no Python:

- **Carregamento dos dados:** Usei a biblioteca `pandas`.
- **Exploração Inicial:** Utilizei `df.info()`.`describe()` para observar a estrutura dos dados e para ver suas estatísticas.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

Tratamento de Valores Ausentes

Para identificar a presença e a quantidade de valores ausentes em cada coluna do Dataset, foi utilizada a função `.isnull().sum()`. Esta função retorna a soma de valores True (que indicam NaN ou null) para cada coluna, fornecendo uma visão clara das lacunas nos dados.

	0
Customer ID	0
Age	0
Gender	0
Item Purchased	0
Category	0
Purchase Amount (USD)	0
Location	0
Size	0
Color	0
Season	0
Review Rating	37
Subscription Status	0
Shipping Type	0
Discount Applied	0
Promo Code Used	0
Previous Purchases	0
Payment Method	0
Frequency of Purchases	0

Tratamento de Dados Ausentes: Verifiquei a presença de valores nulos e os valores faltantes na coluna **Review Rating** que foi preenchido utilizando a mediana das avaliações de cada categoria de produto.

Padronização de Colunas: As colunas foram renomeadas para o formato *snake case* para melhorar a legibilidade e a documentação.

Engenharia de Atributos: Foi criada a coluna **age_group** agrupando as idades dos clientes em faixas etárias, e a coluna **purchase_frequency_days** a partir dos dados de compras.

Verificação de Consistência dos Dados: Foi analisado se **discount_applied** e **promo_code_used** eram redundantes; a coluna **promo_code_used** foi removida.

Tradução: Tradução de todas as colunas e registros para o idioma Portugues para melhor entendimento.

Integração com Banco de Dados: O dataset do script Python foi limpo e baixado para o MySQL Workbench para análise SQL.

Análise de dados usando SQL (KPIs)

Transformei a estrutura de análise MySQL em perguntas de negócio.

1) Qual é a receita total gerada por clientes do sexo masculino vs feminino?

genero	valor
Masculino	157890
Feminino	75191

2) Quais clientes usaram um desconto, mas ainda assim gastaram mais do que o valor médio de compra?

id_cliente	valor_compra
2	64.00
3	73.00
4	90.00
7	85.00
9	97.00
12	68.00
13	72.00
16	81.00
20	90.00
22	62.00
24	88.00

3) Quais são os 5 produtos com a maior avaliação média de avaliações?

item_comprado	Avaliacao do Produto
Luvas	3.86
Sandálias	3.84
Botas	3.82
Chapéu	3.8
Saia	3.78

4) Compare os valores médios de compra entre frete padrão e frete expresso.

tipo_envio	valor
Expresso	60.48
Padrão	58.46

5) Qual é a contribuição de receita de cada faixa etária?

faixa_etaria	valor_total
Jovem	62143
Meia-Idade	59197
Adulto	55978
Idoso	55763

6) Quais são os 3 produtos mais comprados em cada categoria?

produto_placar	categoria	item_comprado	total_clientes
1	Acessórios	Joias	171
2	Acessórios	Cinto	161
3	Acessórios	Óculos de Sol	161
1	Calçados	Sandálias	160
2	Calçados	Sapatos	150
3	Calçados	Tênis	145
1	Roupas	Blusa	171
2	Roupas	Calças	171
3	Roupas	Camisa	169
1	Vestuário Externo	Jaqueta	163
2	Vestuário Externo	Casaco	161

7) Clientes assinantes gastam mais? Compare o gasto médio e a receita total entre assinantes e não assinantes.

assinatura	total_clientes	valor_medio_compra	total
Sim	1053	59.49	62645
Não	2847	59.87	170436

8) Quais 5 produtos têm a maior porcentagem de compras com desconto aplicado?

item_comprado	taxa_desconto
Chapéu	50.00
Tênis	49.66
Casaco	49.07
Suéter	48.17
Calças	47.37

9) Segmente os clientes em Novos, Retornantes e Fiéis com base no total de compras anteriores, e mostre a contagem de cada segmento.

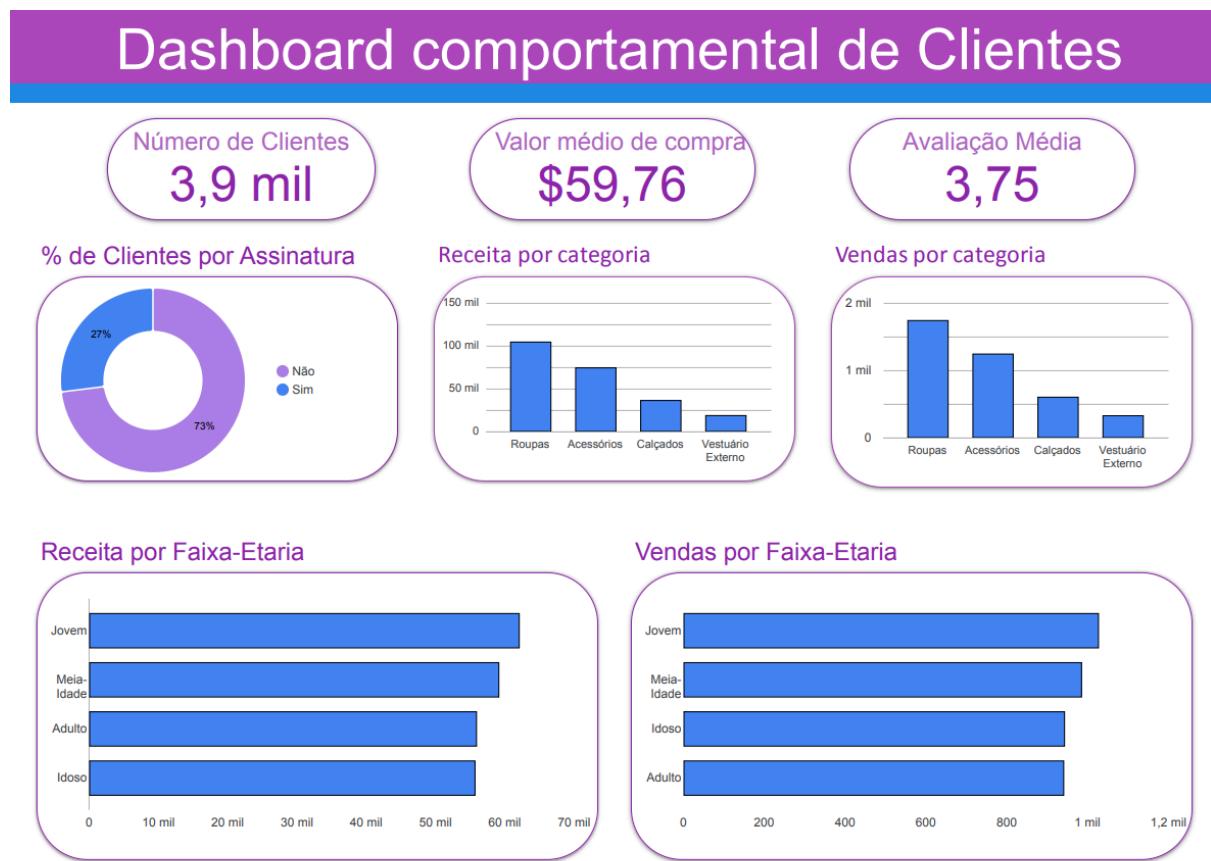
segmento_de_clientes	total_clientes
Fiél	3116
Retornante	701
Novo	83

10) Clientes que compram repetidamente (mais de 5 compras anteriores) também tendem a assinar?

assinatura	compradores_recorrentes
Sim	958
Não	2518

Dashboard no Looker Studio

Por fim, coloquei meu banco de dados no ar utilizando o **Railway**, o que possibilitou conectá-lo ao **Looker Studio** e construir dashboards para a visualização dos insights.



Recomendações de Negócio

Assinaturas Premium: Promova benefícios exclusivos para assinantes.

Programas de Fidelidade: Recompense os compradores recorrentes para moverlos para o segmento “Fiel”.

Política de Descontos nas Avaliações: Equilibre o aumento das vendas com o controle da margem.

Posicionamento de Produtos: Destaque produtos mais bem avaliados e mais vendidos nas campanhas.

Marketing Direcionado: Foque os esforços em faixas etárias de maior receita e em usuários que utilizam envio expresso.