Análise de Dados da plataforma Instacart

Mercado que a plataforma está inserida

O Instacart é uma plataforma de entrega de alimentos para supermercados e compradores. A Instacart ajuda pessoas dos EUA e do Canadá a comprar mais de um bilhão de produtos em mais de mil varejistas com dezenas de milhares de lojas, conectando-os a centenas de milhares de compradores que fazem as compras para eles, 24 horas por dia, onde o cliente pode selecionar digitalmente o supermercado desejado, montar e adquirir sua lista de compras e a receber em casa por meio de um “personal shopper”. Uma base de dados anonimizada foi disponibilizada à comunidade do Kaggle no contexto de uma competição, permitindo a análise do comportamento de compra dos clientes.

Objetivo do Projeto

O objetivo deste projeto foi desenvolver um estudo de caso a partir do dataset Instacart disponível no Kaggle com o objetivo de simular o fluxo de trabalho real de um analista de dados. O foco principal foi a aplicação de técnicas de Data Wrangling (limpeza e união de dados) e Análise Exploratória (EDA) para extrair insights sobre comportamento de consumo, retenção de usuários e performance de categorias de produtos.

Ambiente de Desenvolvimento

O projeto foi desenvolvido em ambiente local utilizando Python e o editor VS Code. Essa escolha permite maior controle do pipeline de dados, versionamento de código e simula práticas adotadas em ambientes corporativos de análise e engenharia de dados.

Entendendo o Negócio e os Dados

O conjunto de dados disponibilizado foi construído pensando no fluxo do cliente no processo de compra. O fluxo de dados foi modelado para representar, de forma fiel, a jornada de compra do cliente dentro da plataforma Instacart.

O processo inicia-se quando um cliente (user_id) acessa a plataforma e realiza um pedido. Cada pedido recebe um identificador único (order_id), que permite rastrear todas as informações relacionadas àquela compra específica.

Dentro de um pedido, o cliente pode adicionar um ou vários produtos, cada um identificado por um product_id e descrito pelo seu nome (product_name). À medida que os produtos são incluídos no carrinho, o sistema registra a ordem de adição

(add_to_cart_order), capturando o comportamento de navegação e priorização do cliente durante a compra.

Além disso, o sistema registra o dia da semana (order_dow) e a hora do pedido (order_hour_of_day), permitindo análises temporais sobre hábitos de consumo.

Para garantir o correto rastreamento e categorização dos produtos, cada item é associado a um corredor (aisle) e a um departamento (department), o que possibilita análises por categoria, identificação de preferências e estudos de recompra por tipo de produto.

Com o objetivo de compreender o comportamento do cliente ao longo do tempo, são registradas informações históricas como:

- os dias desde o último pedido (days_since_prior_order), que ajudam a medir recência, métrica importante para a Matriz RFV (Recência, Frequência e Valor) que é uma técnica de segmentação que classifica clientes com base no comportamento de compra para otimizar estratégias de marketing. Ela identifica quem comprou recentemente (R), com que frequência compra (F) e quanto gasta (V), permitindo ações personalizadas para retenção, fidelização e aumento de vendas.
- o número sequencial do pedido (order_number), que indica a evolução da jornada de compra do cliente.

Por fim, o sistema identifica se um produto comprado já havia sido adquirido anteriormente pelo mesmo cliente, por meio do indicador reordered, permitindo a análise de recompra, fidelidade e padrões de consumo recorrente. O conjunto de dados é composta pelas seguintes tabelas:

1. **orders.csv**

Principais colunas:

- user_id (PK): Identificador do cliente.
- order_id (PK): Identificador único do pedido.
- order_number: Número sequencial do pedido realizado por cada cliente. Iniciando sempre em 1, que representa o primeiro pedido do cliente, e aumentando conforme a ordem cronológica dos pedidos realizados.
- order_dow: Dia da semana em que o pedido foi realizado, codificado de 0 a 6, seguindo uma representação numérica cíclica dos sete dias da semana. Sem especificar exatamente quando começa a semana. Por
- order_hour_of_day: Hora do dia em que o pedido foi realizado, representada em formato inteiro de 0 a 23.

- days_since_prior_order: Dias desde o pedido anterior do mesmo cliente. A coluna é gerada automaticamente pelo sistema a partir do histórico de pedidos de cada cliente. Para cada pedido a partir do segundo, o sistema calcula: Diferença em dias entre a data do pedido atual e a data do pedido imediatamente anterior do mesmo cliente. Esse valor é então armazenado em days_since_prior_order. Atenção: no primeiro pedido não existe pedido anterior para comparação logo o sistema não calcula o intervalo e o campo é registrado como NULL / NaN. O sistema registra 0 (zero) quando o cliente realizou dois pedidos no mesmo dia. Além disso, o registro do intervalo entre pedidos está limitado a um máximo de 30 dias. Valores iguais a 30 representam intervalos de 30 dias ou mais desde o pedido anterior, evitando valores extremos e facilitando análises comportamentais e preditivas.

Vale ressaltar que as principais limitações do dataset estão relacionadas à ausência de informações financeiras, demográficas e temporais completas, e não a problemas de inconsistência ou erro nos dados.
Além disso, a base apresenta algumas particularidades importantes, como a presença de valores NaN, que refletem regras de negócio (por exemplo, no primeiro pedido do cliente).

2. **order_products__prior.csv**

- order_id (PK): Identificador único do pedido
- product_id (FK): Identificador do produto
- add_to_cart_order: Ordem em que o produto foi adicionado ao carrinho por pedido. Começa no 1 e reinicia a cada novo pedido.
- reordered: Indicador se o produto foi recompra (1 = sim, 0 = não). A coluna reordered é um indicador binário que informa se um produto comprado em um pedido já havia sido comprado anteriormente pelo mesmo cliente.

3. **products.csv**

- product_id (PK): Identificador do produto
- product_name: Nome do produto
- aisle_id (FK): Identificador do corredor
- department_id (FK): Identificador do departamento

4. **departments.csv**

- department_id: Identificador do departamento
- department: Nome do departamento

5. **aisles.csv**

- aisle_id (PK): Identificador do corredor
- aisle: Nome do corredor. O corredor é uma subcategoria dentro de um departamento

Preparação dos dados

Importação e Inspeção dos Dados

Os dados foram extraídos por meio da API da plataforma Kaggle, garantindo reprodutibilidade e aderência a boas práticas de ingestão de dados. A ingestão dos dados foi realizada por meio da API oficial do Kaggle, permitindo o download automatizado e reprodutível do dataset Instacart Market Basket Analysis.

Foi realizada uma inspeção inicial nos dados com o objetivo de compreender a estrutura das bases, identificar padrões e verificar a presença de valores nulos.

O notebook [01_ingestao_e_exploracao.ipynb] foi responsável pela ingestão dos dados brutos e pela exploração inicial das bases, com foco na compreensão da estrutura, qualidade e contexto de negócio dos dados.

Na etapa de verificação de dados nulos, a coluna days_since_prior_order apresentou valores ausentes. A análise do contexto dos dados indica que esses valores não representam falhas de coleta, mas sim uma condição esperada do negócio.

Essa variável representa a quantidade de dias desde o pedido anterior do cliente. Portanto, nos casos em que o pedido analisado corresponde ao primeiro pedido do cliente, não existe um pedido anterior para o cálculo do intervalo, fazendo com que o sistema registre esse valor como nulo. Neste momento do projeto, essa informação foi mantida dessa forma.

Lidar com Dados Ausentes e Integração das tabelas

O notebook [02_preparacao_dados_etl.ipynb] teve como objetivo realizar a preparação dos dados e a integração das tabelas.

Por fim, foi executado o processo de integração de múltiplas tabelas relacionais, resultando na construção da tabela base_compras_tratada, que consolida os eventos de compra no nível mais granular e salva no ambiente local.

Explorando as Características dos Dados

O notebook [03_analise_descritiva_base_integrada] teve como objetivo realizar a análise exploratória da base que contém 206.209 clientes únicos, o que representa uma amostra estatística expressiva para a identificação de padrões de consumo, hábitos de compra e comportamento recorrente. Foram identificados 3.214.874 pedidos únicos. Além disso, o catálogo de produtos conta com 49.677 itens únicos.

Estudo de caso

Este estudo de caso foi estruturado como um cenário simulado, visando reproduzir os desafios analíticos e as demandas estratégicas do cotidiano de um Analista de Dados.

1. O Desafio

A equipe de Marketing do Instacart disponibilizou uma base de dados transacionais abrangente com o objetivo estratégico de decifrar o comportamento do consumidor. O desafio central consiste em transformar dados brutos em inteligência de mercado para responder a três perguntas fundamentais:

- Quando compram? (Sazonalidade diária e horária).
- Como compram? (Jornada do cliente, tamanho da cesta e hábitos de recompra).
- O que os clientes compram? (Mix de produtos e afinidade de categorias).

2. Análise Temporal

A variável days_since_prior_order representa o intervalo de tempo, medido em dias, entre o momento em que um usuário finalizou um pedido e o momento em que ele fez o pedido imediatamente anterior.

A figura 1 revela que 50% dos clientes realizam novos pedidos em até 8 dias (mediana), indicando um ciclo de retorno rápido. Esse comportamento é fortemente impulsionado pelo hábito semanal, como demonstra o pico expressivo no 7º dia, que representa a moda do conjunto de dados. A distribuição apresenta uma assimetria positiva, confirmada pelo fato de a média (11,10 dias) ser superior à mediana. Visualmente, esse fenômeno é validado pela maior concentração de pedidos nos primeiros dias (à esquerda) e por uma cauda longa à direita, influenciada por clientes com intervalos de recompra extensos ou pelo acúmulo de registros no limite de 30 dias.

Relevância para a análise:

- Ciclos semanal de compra com picos no 7 dia.
- Recompra com período extenso acima de 15 dias

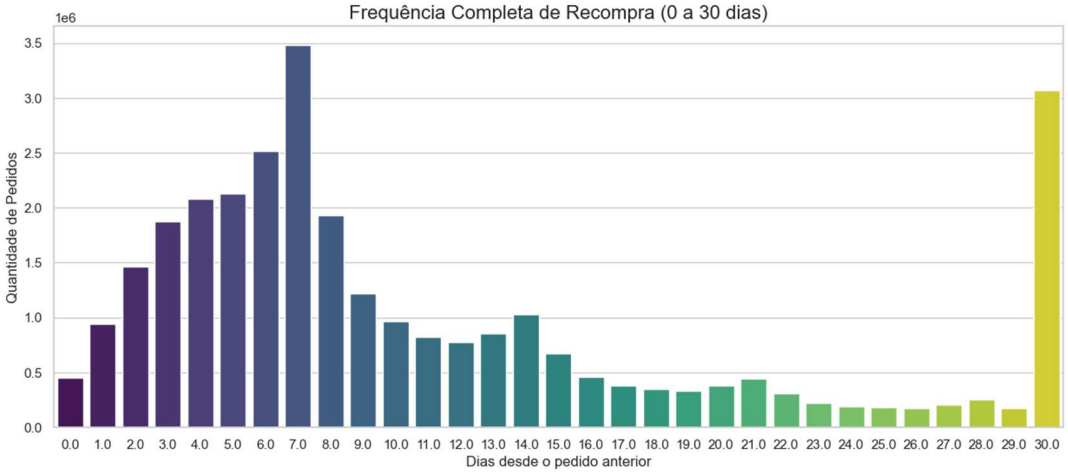


Figura1. Gráfico frequência x quantidade de pedido
Fonte: Resultados originais da análise

A análise da variável order_dow identifica o dia da semana em que cada transação foi realizada. Os dados foram codificados em um intervalo de 0 a 6, seguindo o padrão internacional de indexação onde o valor 0 é atribuído ao domingo. Essa padronização permite mapear o comportamento de compra ao longo do ciclo semanal.

A Figura 2 evidencia uma sazonalidade semanal marcada pela elevada volumetria de transações entre domingo e segunda-feira, sugerindo um comportamento de compra voltado ao abastecimento doméstico para o início da semana.

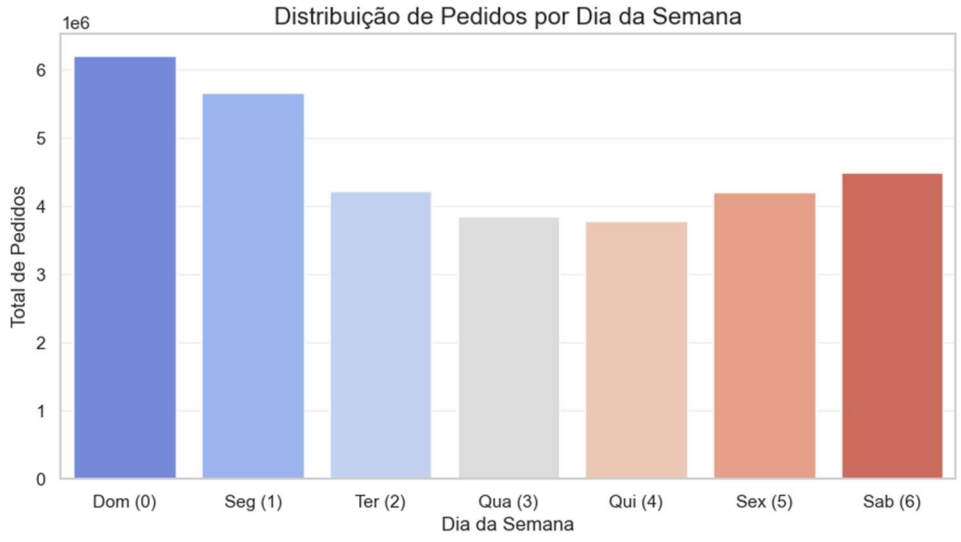


Figura 2. Distribuição de pedidos por dia da semana
Fonte: Resultados originais da análise

A Figura 3 identifica uma concentração massiva de pedidos (representada pelas áreas em azul escuro) em dois momentos principais: domingo (0) e segunda-feira (1). O pico de acessos no domingo ocorre entre as 10h e 15h, enquanto na segunda-feira o volume se mantém elevado das 9h até as 11h. Observa-se que a janela de compras inicia sua ascensão diária às 7h e atinge o ápice de conversão entre 10h e 16h.

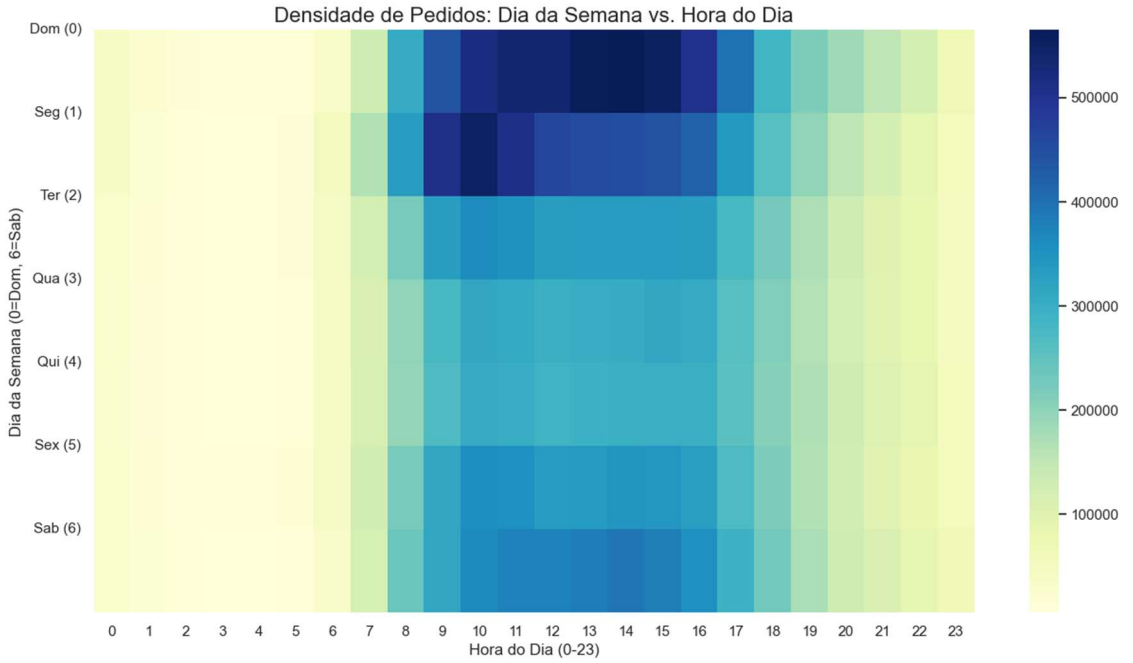


Figura 3. Mapa de Calor densidade de pedidos
Fonte: Resultados originais da análise

3. Frequência de pedidos

A variável order_number registra a sequência cronológica dos pedidos por usuário. Ao aplicar a função de agregação max() agrupada por cliente, foi possível determinar a frequência total de compra de cada indivíduo.

A análise estatística da variável total_pedidos (extraída através do valor máximo de order_number por cliente) revela uma distribuição com forte assimetria à direita (positiva). Esta configuração é evidenciada pelo distanciamento entre as medidas de tendência central, onde a média (16,23) supera significativamente a mediana (10,0) e a moda (4,0). A figura 4 demostra uma concentração massiva de usuários está no intervalo entre 4 e 10 pedidos, conforme observado no pico do histograma. Em contrapartida, a média elevada é sustentada por uma “cauda longa” de usuários fidelizados (outliers positivos), que realizam um volume de pedidos muito acima da norma populacional, chegando ao limite registrado de 100 transações.

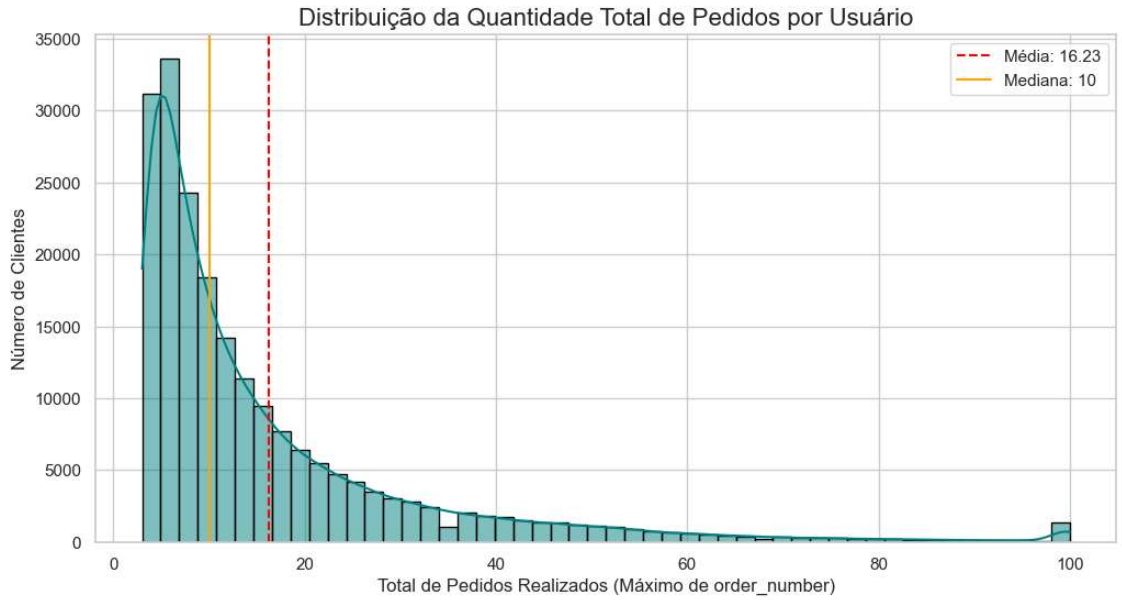


Figura 4. Distribuição da quantidade total de pedidos por usuário
Fonte: Resultados originais da análise

4. Quantidade de itens por usuário

O volume de itens permite calcular o "Ticket Médio" em volume de itens, identificando se os clientes fazem compras de abastecimento (muitos itens) ou de conveniência (poucos itens).

A figura 7 revela o comportamento típico da cesta de compras do usuário. A assimetria à esquerda confirma que a maioria das transações é composta por poucos itens com mediana de 8 itens.

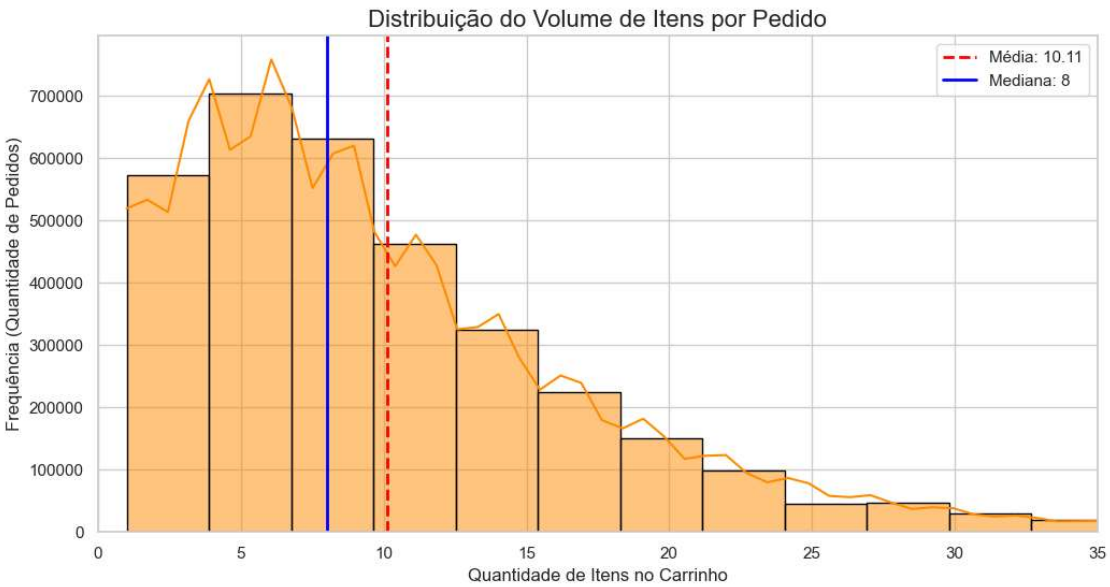


Figura 7. Distribuição do volume de itens por pedido
Fonte: Resultados originais da análise

5. Segmentação de Produtos

A figura 5 mostra a arquitetura do catálogo analisado que é composta por 21 departamentos e 134 corredores (aisles), apresentando uma ampla capilaridade de produtos. A análise quantitativa do volume de vendas revela uma concentração expressiva em quatro pilares fundamentais: Produce (hortifrúti), Dairy Eggs (laticínios e ovos), Snacks e Beverages (bebidas).

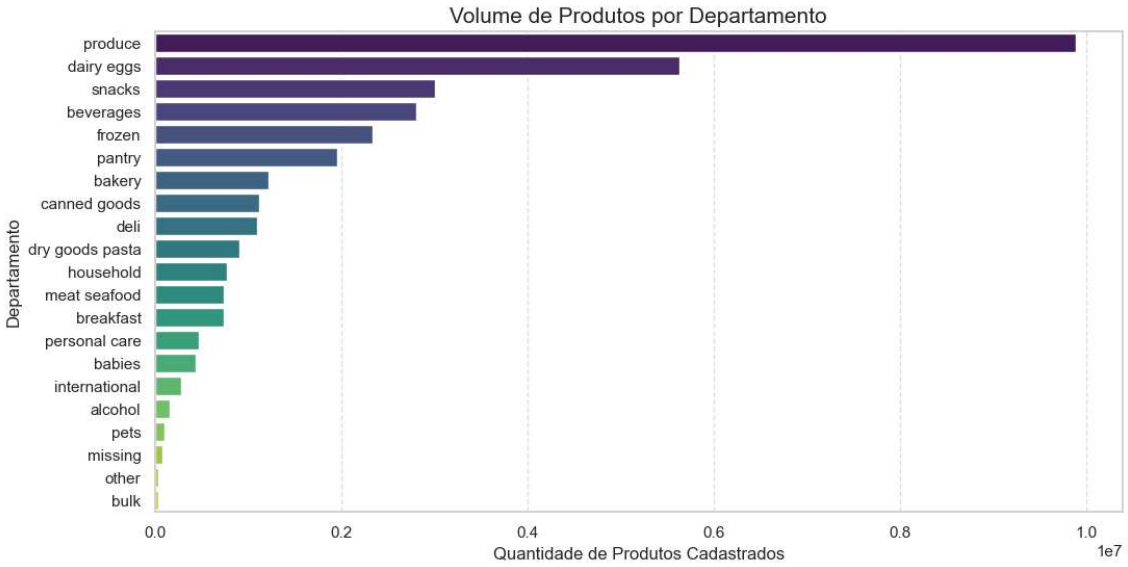


Figura 5. Volume de produtos por departamento
Fonte: Resultados originais da análise

6. Produtos abre carrinho

A figura 6 apresenta os 20 produtos com maior frequência de adição inicial ao carrinho (add_to_cart_order = 1). Estes itens são classificados como “Produtos de Destino’’, pois representam a necessidade primária que motiva o usuário a abrir o aplicativo.

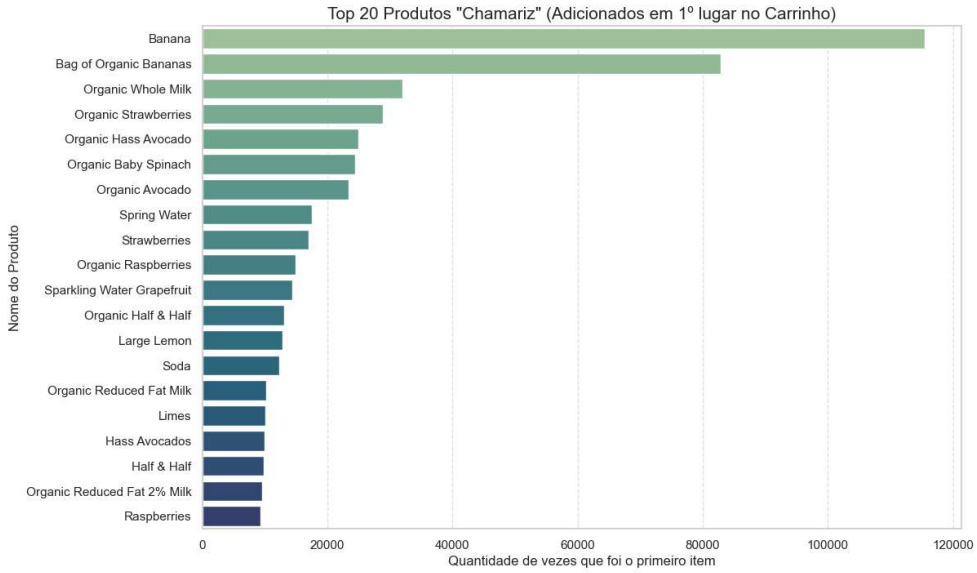


Figura 6. Top 20 de produtos adicionados em 1º lugar no carrinho
Fonte: Resultados originais da análise

7. Recompra

A Figura 8 identifica a hierarquia de fidelização do catálogo, destacando os departamentos com maior tração de recompra. O topo do ranking é ocupado por Dairy Eggs, Beverages e Produce, categorias que apresentam as taxas mais elevadas de recorrência. A Taxa Geral de Recompra consolida-se em 59%. Esse indicador reflete o hábito de consumo recorrente dos usuários na plataforma.

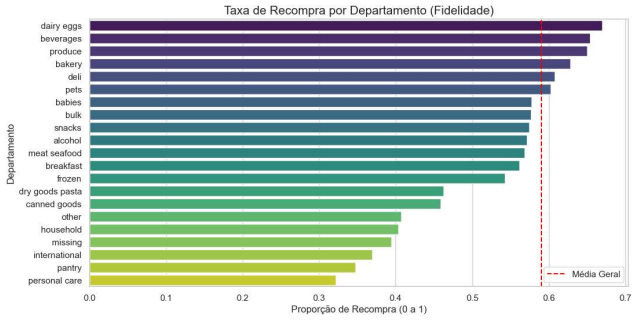


Figura 8. Taxa de recompra por departamento
Fonte: Resultados originais da análise

Dando continuidade à simulação, os resultados agora fundamentam o desenvolvimento do relatório final. Esta etapa é dedicada à tradução das análises técnicas em um formato executivo, destacando os padrões de consumo.

Construção do relatório para a equipe de marketing

A Figura 9 apresenta o dashboard consolidado com os resultados das análises. O comportamento de consumo revela que o sucesso da operação depende da reposição semanal de itens frescos (Produce), que funcionam como o principal gatilho para a abertura do app. O desafio estratégico identificado consiste na conversão de clientes casuais em recorrentes, focando na superação do gargalo de 4 a 10 pedidos por usuário. Para isso, recomenda-se a intensificação de campanhas nos períodos de maior tráfego (domingos e segundas-feiras).

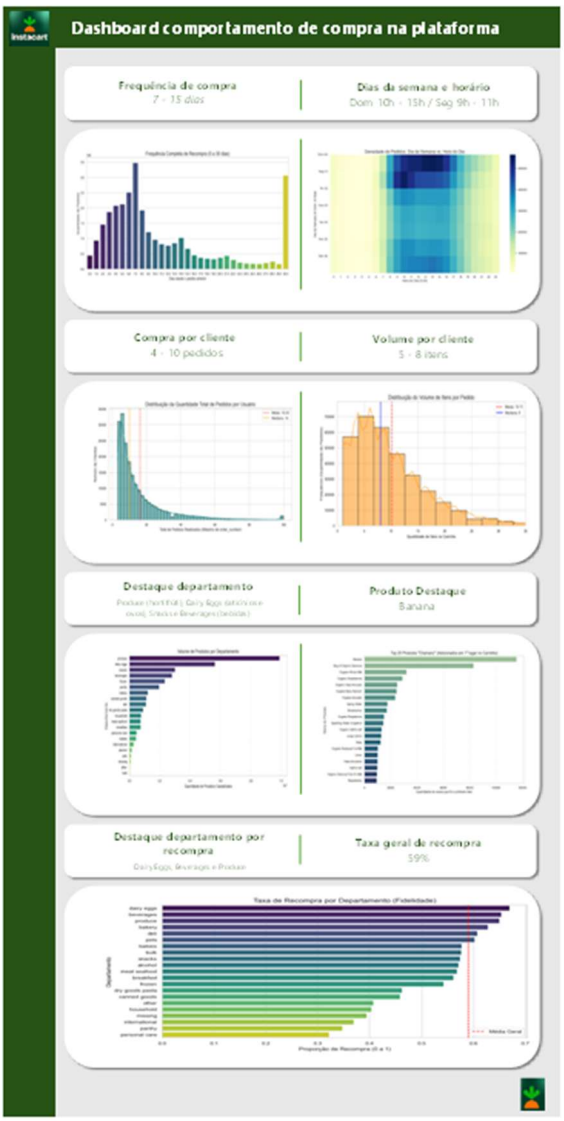


Figura 9. Dashboard
Fonte: Resultados originais da análise

Após a apresentação do dashboard, a equipe de Marketing do Instacart obteve uma visão clara dos gargalos de retenção e dos padrões de compra. Com base nos dados expostos, especialmente a "Barreira dos 4 pedidos" e a dominância dos itens de hortifrúti como porta de entrada; foram levantadas hipóteses para direcionar os próximos passos.

Foi identificado que a Banana atua como o principal de entrada. Para potencializar o faturamento, a equipe de Marketing solicitou uma Análise de Afinidade para mapear o comportamento de compra cruzada. O objetivo é descobrir quais itens possuem maior “aderência’’ à cesta quando a banana está presente, permitindo a criação de sistemas de recomendação mais assertivos e combos promocionais inteligentes.

A técnica Market Basket Analysis identifica padrões de co-ocorrência em transações. No contexto deste estudo, ela revela quais produtos possuem afinidade quando a banana está no carrinho. O notebook [04_outras_analises] teve como objetivo central a extração de padrões comportamentais para fundamentar estratégias de recomendação personalizadas.

Estratégias de Recomendação

A Dualidade da Banana

Como o item mais vendido da plataforma, a Banana permite a implementação de duas lógicas distintas de recomendação no sistema:

1. Cross-Selling (Banana como antecedente)

Neste cenário, o cliente já possui a Banana no carrinho. O sistema utiliza o Lift para sugerir produtos que possuem afinidade real, mas menor volume de venda por si só (como a *Maçã Fuji* ou *Hass Avocado*).

- Objetivo: Aumentar o Ticket Médio (o valor total da compra).
- Lógica: Já que você leva o básico, que tal este item premium que combina com sua cesta?"

2. Lembrete de Relevância (Banana como consequente)

Aqui, o algoritmo atua como um assistente inteligente. Se o cliente coloca itens com alto Lift (como *Limes* ou *Organic Raspberries*) mas ainda não adicionou a Banana, o sistema dispara um lembrete.

- Objetivo: Aumentar a Satisfação e Conveniência (evitar que o cliente esqueça o essencial).

- Lógica: Baseado nos itens frescos que você escolheu, notamos que você ainda não adicionou Bananas. Gostaria de incluí-las?

A análise foca na abordagem da Banana como consequente. Por meio do algoritmo Apriori, identificamos associações entre os 100 produtos mais frequentes, varrendo a matriz transacional para isolar co-ocorrências estatisticamente relevantes.

A etapa final da modelagem consistiu na derivação de regras de associação priorizando a métrica de Lift. Ao estabelecermos um limiar mínimo de 1.0, eliminamos ruídos estatísticos e focamos em relações de interdependência positiva. A ordenação decrescente do Lift permitiu identificar as oportunidades de *cross-selling* de maior impacto, revelando padrões que não seriam visíveis apenas com análises de volume, transformando correlações matemáticas em estratégias de recomendação acionáveis.

O Lift atua como o validador da inteligência do sistema: ele "limpa'' o viés de popularidade da Banana (que possui 20% de suporte), isolando apenas as combinações onde a presença do antecedente realmente alavanca a venda do segundo item. Um exemplo prático é a Maçã Fuji, que gera uma propensão de compra de Banana 88% superior à média da plataforma. Enquanto a Confiança indica a frequência da combinação, o Lift confirma se existe uma conexão real de consumo, ou seja, se um produto de fato atrai o outro ou se a co-ocorrência é meramente casual.

A tabela 1 apresenta os pares de produtos que possuem a maior força de associação com a categoria de bananas, priorizando a métrica de Lift.

Tabela 1: Associações de Alta Afinidade (Âncora: Banana)

Antecedente (Item A)	**Consequente (Item B)**	**Suporte**	**Confiança**	**Lift**
Organic Fuji Apple	**Banana**	1,43%	37,84%	**1,88**
Honeycrisp Apple	**Banana**	1,21%	35,57%	**1,77**
Cucumber Kirby	**Banana**	1,34%	32,92%	**1,64**
Organic Raspberries	**Bag of Org. Bananas**	1,72%	29,65%	**1,83**
Organic Hass Avocado	**Bag of Org. Bananas**	2,64%	29,31%	**1,81**

Fonte: Resultados originais da análise

A análise identifica que a estratégia de *cross-selling* mais eficiente para a Banana não deve ser genérica, mas sim direcionada aos consumidores de frutas específicas (Maçãs e Abacates), onde a conexão de consumo é comprovadamente mais forte.

Após a consolidação dos dados de entrada, a equipe de Marketing identificou o departamento de Hortifrúti (Produce) como a principal porta de entrada e o maior driver de volume da plataforma. O desafio proposto foi: *Uma vez que o cliente garante seus itens frescos, qual é o próximo passo na sua jornada de compra?*

Para responder a esse questionamento, aplicamos uma técnica de filtragem por exclusão. Isolamos todos os carrinhos que continham itens de Hortifrúti e removemos estatisticamente esse departamento da base de cálculo. Esse processo permitiu enxergar o "Segundo Caminho" do cliente, eliminando o viés de auto-correlação.

Análise de Afinidade de Destino

Para responder tecnicamente à demanda da equipe de marketing, implementamos um algoritmo de filtragem que isola a jornada do cliente pós-Hortifrúti identificando pedidos que contêm pelo menos 1 produto do departamento “produce’’. Ao

utilizarmos o método .isin() para selecionar pedidos de referência e, em seguida, aplicarmos uma exclusão categórica do departamento “produce’’, conseguimos mapear o fluxo real de transição.

Tabela 2: O Segundo Caminho do Cliente

Departamento	**Frequência Relativa**
Dairy Eggs (Laticínios)	**24,60%**
Snacks (Lanches)	**12,13%**
Beverages (Bebidas)	**10,47%**
Frozen (Congelados)	**9,77%**

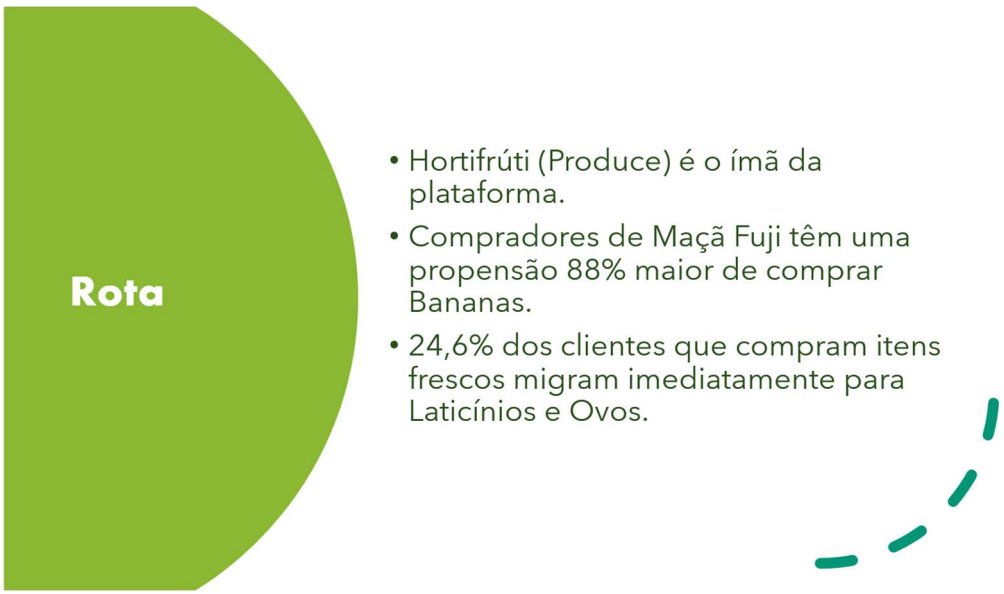
Fonte: Resultados originais da análise

O resultado revela uma dependência crítica do setor de Laticínios e Ovos (24,6%), consolidando-o como o destino prioritário do cliente após Hortifrúti.

Para trabalhos futuros, vale destacar a limitação do algoritmo Apriori em tratar a transação como um conjunto de itens simultâneos, ignorando a cronologia da escolha. Recomenda-se a exploração da métrica add_to_cart_order, que possibilitará o mapeamento do funil de decisão em tempo real. Esta abordagem permitirá transitar de uma análise de “o que foi comprado’’ para “como foi comprado’’, identificando os gatilhos iniciais de consumo e os momentos de maior suscetibilidade.

Apresentação para equipe de marketing

Slides:



Plano de Validação

Otimização do Ticket Médio (Upselling)

- Hipótese: a recomendação de Bananas para quem adiciona Maçãs (alto Lift) reduz a fricção de compra e aumenta o número de itens por cesta.
- Ação (Teste A/B):

1. Grupo A (Controle): recebe recomendações genéricas (ex: "Mais vendidos").
2. Grupo B (Variante): recebe recomendação direta de Bananas logo após adicionar a Maçã.
3. Métrica de Sucesso: aumento na taxa de conversão do item recomendado.

Plano de Validação

Conversão de Cesta

Hipótese: clientes que iniciam a compra pelo Hortifrúti têm uma intenção de abastecimento básico. Sugerir Laticínios precocemente aumenta a probabilidade de fechamento da cesta completa na nossa plataforma.

Ação (Teste A/B):

1. Grupo A (Controle): navegação livre entre departamentos.
2. Grupo B (Variante): exibição de um banner ou "Atalho de Categoria" para Laticínios e Ovos assim que o segundo item de Hortifrúti é adicionado.
3. Métrica de Sucesso: aumento na frequência de pedidos contendo as duas categorias