Linx - Software House

Análise de dados

Vinícius Camozzato Vaz

Araranguá

2020

1. INTRODUÇÃO

Este relatório foi desenvolvido com objetivo de apresentar análises referentes ao conjunto de dados de uma amostra de vendas disponibilizados pela empresa Linx. Estes dados são referentes ao mês de agosto de 2018 de uma empresa que possui diversas lojas físicas espalhadas pelo Brasil e também uma loja virtual. No decorrer do relatório serão apresentados métodos, procedimentos e raciocínios utilizados para extrair informações relevantes do conjunto de dados disponível.

Para realização das análises foi utilizada basicamente a linguagem de programação Python para todas etapas do processo de produção. Sendo as principais ferramentas utilizadas para:

- **Extração e manipulação de dados gerais –** Bibliotecas como Pandas e Numpy.
- **Visualização dos dados –** Bibiliotecas como Matplotlib, Seaborn.
- **Extração e manipulação de grandes volumes de dados –** Apache Spark (PySpark).

2. DADOS

Foram fornecidos três tipos de dados: vendas de lojas físicas, páginas visitadas do ecommerce e pedidos realizados no ecommerce.

2.1 Dados de vendas offline


```
{
   "date": "2018-08-01",
   "state": "RN",
   "store_id": "3162633",
   "sale_id": "666639323036376",
   "off_product_id": "643839313363376",
   "quantity": 1,
   "price": 149.0,
   "customer_id": "30373934343338363136"
}
```

2.2 Dados de vendas online


```
{
   "date": "2018-08-01",
   "visitor_id": "34303165316239643163326l3",
   "deviceType": "mobile",
   "order_id": "356664366366353",
   "on_product_id": "313562333039323",
   "quantity": 1,
   "price": 629.0,
   "customer_id": "63393337303931353431"
}
```


2.3 Dados de páginas visualizadas


```
{
   "date": "2018-08-01",
   "visitor_id": "38326365313735383731373373",
   "deviceType": "mobile",
   "pageType": "product",
   "category_id": "6365313034",
   "on_product_id": "323239323839626",
   "customer_id": "333431633165643132 64"
}
```


Dos dados apresentados, o **customer_id** é o mesmo para todos conjuntos porém um valor de **on_product_id** não representa o mesmo produto para este valor em **off_product_id**.

3.0 MÉTRICAS E QUESTÕES

Foi solicitado pela empresa que fossem respondidas algumas perguntas, sendo elas:

1) **Qual foi o faturamento total no período?**
2) **Qual o produto mais comprado online?**
3) **Cariocas gostam de comprar no fim de semana?**
4) **É comum escolher online e terminar a compra na loja física?**
5) **O time de marketing desta rede quer fazer uma campanha oferecendo um cupom de 20% nas compras de loja física para quem visitou o site e abandonou um carrinho com produtos. Estime o resultado dessa campanha.**

Para realização da análise foi adotado dois tipos de métricas, a primária que condiz com a métrica principal do projeto e testada em produção e a secundária que não foi testada pois corresponde as métricas a serem monitoradas na etapa de pós-produção, o qual não haviam dados disponíveis no momento deste relatório.

3.1 PRIMÁRIAS

Nas métricas primárias estão presentes indagações referentes às questões solicitadas pela empresa, além destas, também estão presentes questões possivelmente relevantes a serem respondidas durante a análise.

- **Período de coleta**
- **Faturamento total**
- **Faturamento total Lojas Físicas**
- **Faturamento total Online**
- **Loja Física com maior faturamento no mês**
- **Loja Física com menor faturamento no mês**
- **Média de faturamento por dia dos estados com maior arrecadação**
- **Faturamento total de todas por dia da semana (Online e Offline)**
- **Faturamento total por dia do mês (Online e Offline)**
- **Faturamento total por estado**
- **Faturamento total por dia sobre dispositivo utilizado para compra**
- **Número total de vendas no período**
- **Dias com maior arrecadação (Online e Offline)**
- **Produto mais visualizado na Loja Virtual**
- **Produto mais comprado na Loja Virtual**
- **Quantidade de unidades dos principais produtos vendidos online por dispositivo de acesso.**
- **Faturamento dos principais produtos vendidos online por dispositivo de acesso.**
- **Representatividade das vendas no final de semana (RJ)**
- **Representatividade do faturamento no final de semana (RJ)**
- **Número de clientes identificados que utilizam ambos serviços**
- **Número de visitantes que abandonam carrinho com produto**
- **Quantidade de clientes identificados que abandonam produto**
- **Quantidade de usuários identificados nas páginas visitadas por dispositivo**
- **Quantidade de visitas com abandono de carrinho por dispositivo**

3.2 SECUNDÁRIAS

As métricas sugeridas aqui são feitas por curiosidade e não foram testadas no momento desta produção. Elas seriam referentes principalmente à feedbacks gerados a partir das ações tomadas na etapa de produção, como por exemplo do time de marketing. Considerando que o time de marketing fez alguma campanha que gerasse desconto nos produtos, poderiam ser extraídos dados para verificação da eficácia da pós-produção como:

- **Número de devoluções de clientes que ganharam desconto**
- **Número de trocas de produtos desses clientes**
- **Quantas pessoas que receberam desconto voltaram a comprar na loja após o cupom (médio prazo)**

4.0 ANÁLISES

Para realização das análises foram utilizadas essencialmente as ferramentas descritas na seção **1**. Começamos as análises pelo conjunto de dados das compras nas lojas físicas, posteriormente utilizamos os dados das compras na loja virtual então passando para os dados de páginas visualizadas. Além de análises isoladas sobre os 3 conjuntos de dados, foram realizadas análises combinadas sobre os conjuntos disponíveis para extração da maior quantidade de informações possível.

Precedentemente a etapa de extração concreta de dados de valor, passamos por duas etapas de muita importância na análise dos dados, sendo elas a **Análise Exploratória dos Dados** que tem como objetivo principal entender melhor o conjunto de dados a ser trabalho e sua capacidade de fornecer informação e a **Preparação e Tratamento dos Dados** que foca na limpeza dos dados. Essa etapa contou com duas hipóteses de coleta, uma vez que a origem dos dados e método de coleta eram desconhecidos. Essas hipóteses são:

- **Dados fornecidos são encontrados em produção**
- **Dados coletados não estão enviesados por seleção**

Além das hipóteses descritas acima, a etapa foi resumida basicamente em 4 passos:

- **Tratamento de tipos**
- **Verificação e tratamento de dados faltantes**
- **Verificação e tratamento de outliers**
- **Verificação e tratamento de dados duplicados exatos**

Após o tratamento dos dados podemos iniciar as análises e extração de informações.

4.1 LOJAS FÍSICAS

O conjunto de dados contém um total de 29372 linhas e as 8 colunas mostradas na seção **2.1**. Como é um conjunto de tamanho pequeno, podemos trabalhar diretamente com Pandas carregando todos os dados sem haver problema de memória.

Notamos diversos dados faltantes da coluna **customer_id** no conjunto de dados das lojas físicas. Adotamos como hipótese que esses dados faltantes são devidos à algum problema na coleta apenas do ID do cliente e que isso não influencia na confiabilidade dos outros dados referentes aos IDs faltantes.

O faturamento total nas lojas físicas neste período foi de **R$ 13747940.13**.

Fizemos observações referentes aos diversos estados onde existem lojas desta empresa.

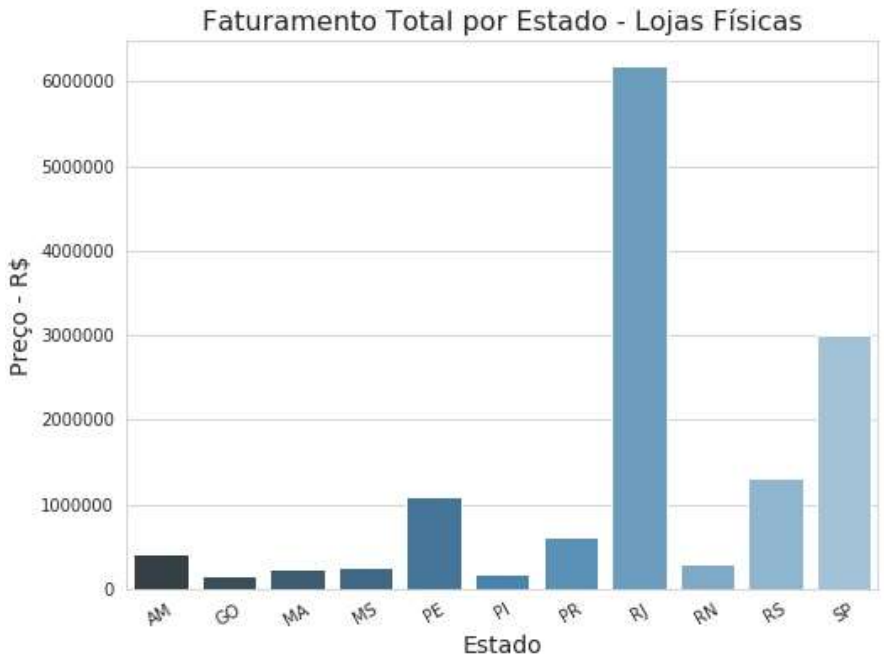

Figura 1

Nota-se que o estado do Rio de Janeiro é o estado que tem maior faturamento, seguido de São Paulo, Rio Grande do Sul e Pernambuco. Isso faz sentido observando a tabela abaixo que representa o número de lojas físicas por estado.

Estado	Número de Lojas
RJ	13
SP	10
RS	3
PE	3
PR	3
AM	2
PI	1
RN	1
MA	1
GO	1
MS	1

Tabela 1

Também se nota que o estado do Paraná, mesmo tendo a mesma quantidade de lojas de Pernambuco e Rio Grande do Sul teve quase metade do faturamento desses estados.

Observamos os 4 estados com maior faturamento para comparar os diferentes comportamentos dos gastos dos clientes sobre cada dia da semana nestes estados que, representam 3 regiões do Brasil distintas (Sul, Sudeste, Nordeste).

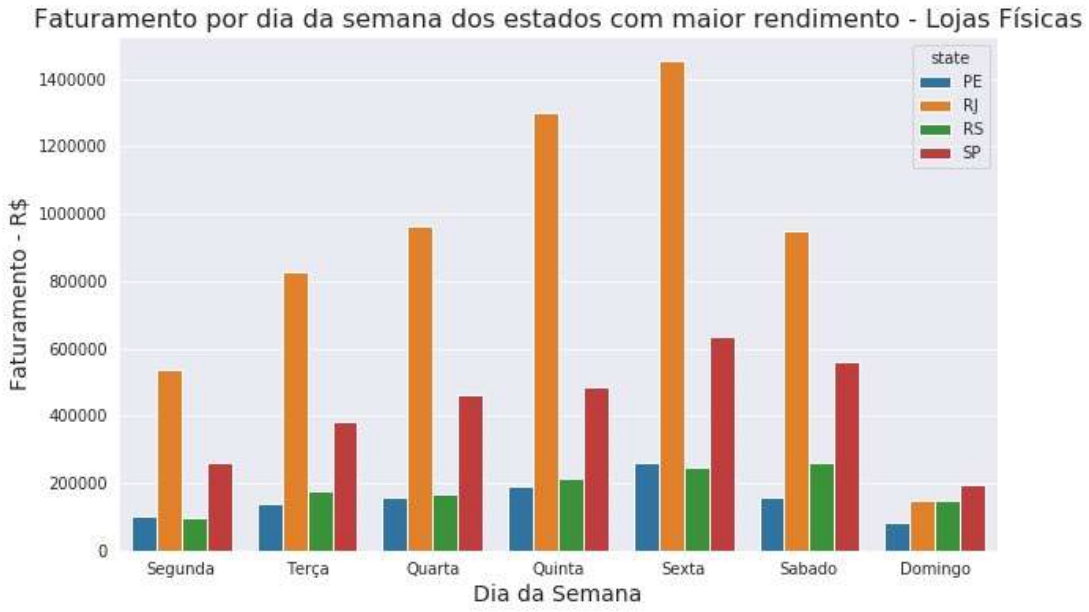

Figura 2

O faturamento de cada dia da semana deste gráfico representa o somatório dos faturamentos do respectivo dia durante o mês, ou seja, para uma segunda-feira por exemplo, o gráfico expressa o somatório do faturamento nas segundas-feiras existentes naquele mês. A tabela abaixo nos provê a informação da frequência de cada dia da semana no mês analisado, com isso aplicamos essa informação ao gráfico acima para obter a média do faturamento dos respectivos dias.

Dia da Semana	Quantidade no mês
Domingo	4
Segunda	4
Terça	4
Quarta	5
Quinta	5
Sexta	5
Sábado	4

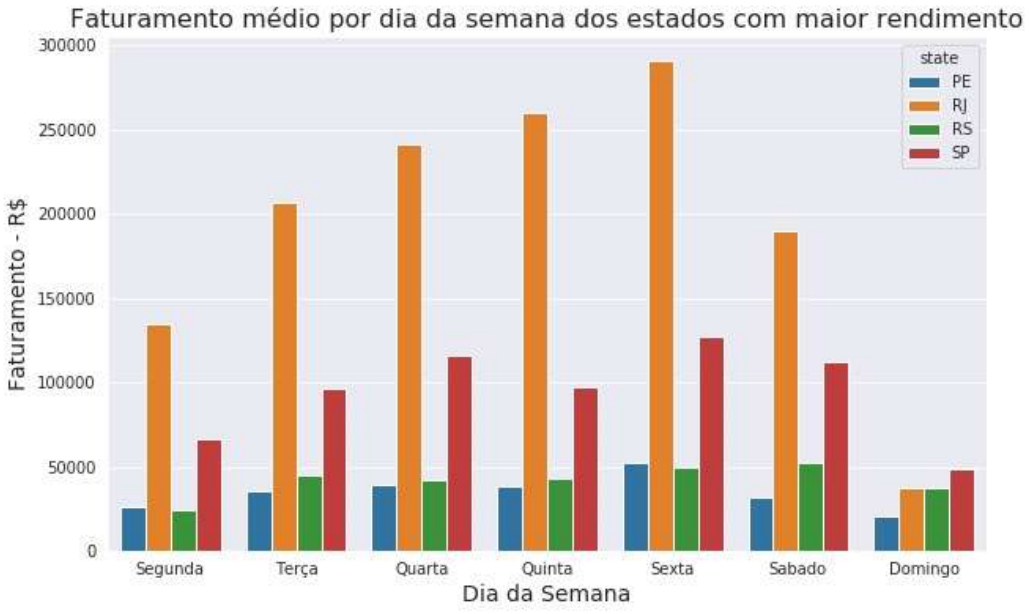

Figura 3

Observando o gráfico podemos notar que a maior tendência de gastos em compras se dá, de maneira geral, nas sextas feiras. Sendo que os Cariocas tendem a gastar mais durante a semana do que sábado e domingo. No domingo, especificamente notamos uma baixa muito grande no faturamento, analisando a quantidade de registro de lojas nos dias de semana, no sábado e no domingo obtivemos:

Dia da semana	Número de registros de lojas (Total)
Dias da semana	39
Sábado	39
Domingo	34

Notamos a falta de registros de algumas lojas referentes à domingos, podem existir lojas que não abrem neste dia, causando baixa na arrecadação.

Figura 4

Analisando o gráfico notamos que o dia com faturamento total máximo foi o último dia do mês.

Além das informações mostradas nos gráficos já apresentados, extraímos os seguintes dados contidos na tabela abaixo:

Dados - Lojas Físicas	Valores
Faturamento Total Lojas Físicas (R$)	13747940.13
Número total de vendas no período (qtd)	14619,00
Faturamento total nos finais de semana (R$)	3092634.48
Faturamento total nos dias de semana (R$)	10655305.65
Loja com maior faturamento no mês (ID-Estado)	6361373-RJ
Loja com menor faturamento no mês (ID-Estado)	3234666-RJ
Dia com maior arrecadação	31/8/2018 - Sexta
Número total de vendas no final de semana (RJ)	887
Número total de vendas no dias de semana (RJ)	3514
Total de vendas no mês (RJ)	4401
Representatividade das vendas no final de semana (RJ)	20.15%
Faturamento no final de semana (RJ)	1097666.58
Faturamento total (RJ)	6179551.72
Representatividade do faturamento no final de semana(RJ)	17.76%

Tabela 2

4.1 LOJA VIRTUAL

O conjunto de dados contém um total de 12237 linhas e as 8 colunas mostradas na seção **2.1**. Como é um conjunto de tamanho pequeno, podemos trabalhar diretamente com Pandas carregando todos os dados sem haver problema de memória.

Notamos também diversos dados faltantes da coluna **customer_id** no conjunto de dados da loja virtual. Adotamos como hipótese que esses dados faltantes são devidos à algum problema na coleta apenas do ID do cliente e que isso não influencia na confiabilidade dos outros dados referentes aos IDs faltantes. Esses valores nulos representam 36% do total dos dados, é um valor significantemente alto presente na falta de identificação.

Além da quantidade de valores nulos, ao analisar o preço dos produtos comprados notamos que existiam alguns produtos de valores muito grande fazendo com que a média da distribuição fosse deslocada para cima, adotamos a mediana para melhor representatividade da medida média de gastos dos usuários.

Extraímos dos dados os 5 produtos mais vendidos online em quantidade.

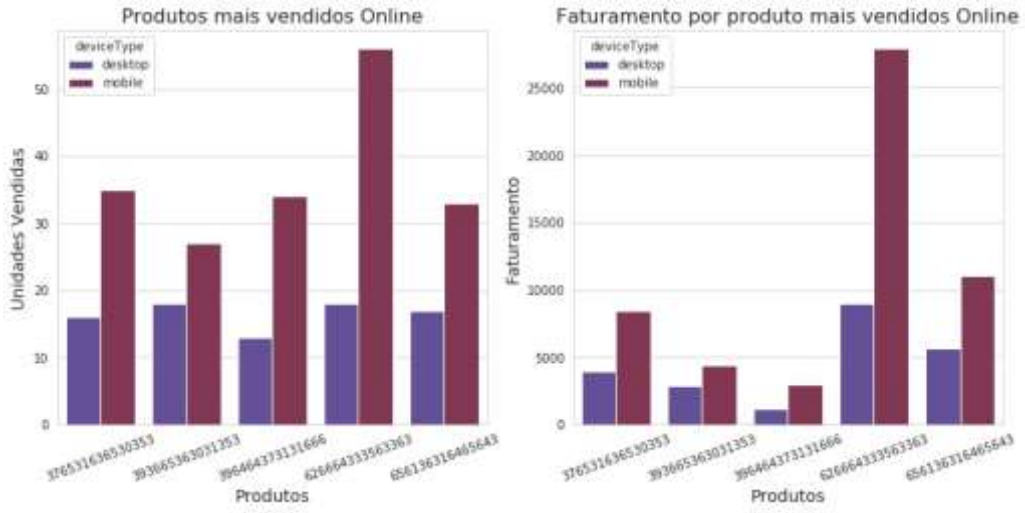

Figura 5

Notamos a partir dos produtos mais comprados online que as pessoas, no geral, tendem a comprar utilizando dispositivos da categoria **mobile**.

O produto com maior número de vendas no período foi o que apresenta a identificação **626664333563363**, e as vendas deste produto ocorreram durante apenas uma semana do mês como mostra a Figura 6.

Figura 6

No restante do período não ocorreram vendas deste produto, possivelmente tenha acabado o estoque e não houve reposição.

O faturamento total da loja virtual no período foi de **R$ 5641837,00**.

A tendência do faturamento pelos dias do mês é mostrada abaixo:

Figura 7

No meio do mês temos um pico de gastos em compras de aproximadamente 5 dias onde o faturamento cresce muito e no restante do mês os valores se mantém próximos.

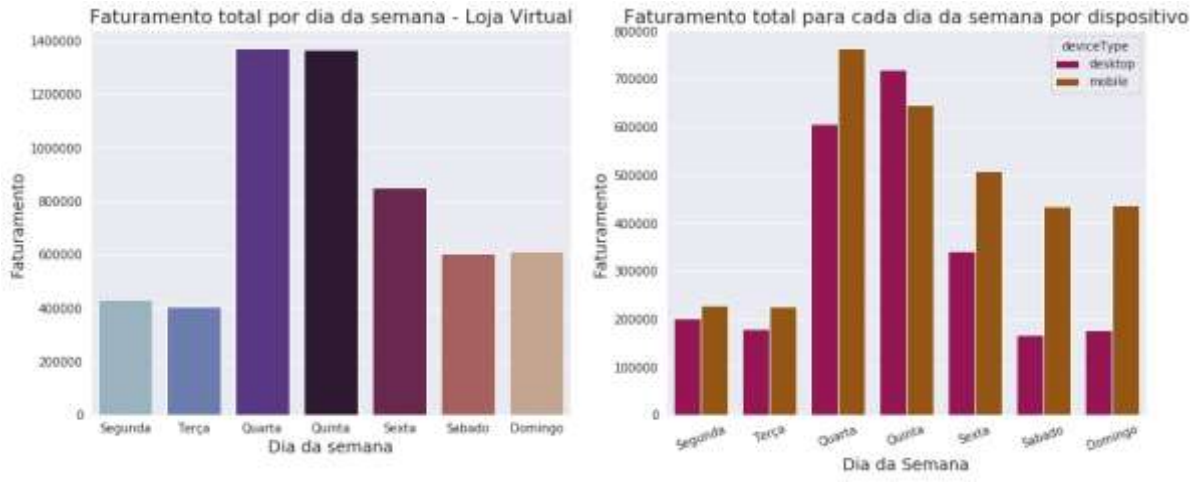

Figura 8

O somatório dos gastos nas quintas e quartas feiras foram os maiores do período, e apenas na quinta feira o gasto total utilizando desktop foi superior ao gasto gerado por compras em dispositivos mobile.

Além das informações mostradas nos gráficos já apresentados, extraímos os seguintes dados contidos na tabela abaixo:

Dados - Loja Virtual	Valores
Faturamento Total Online	5641837
Faturamento finais de semana (R$)	1218498
Faturamento dias de semana (R$)	4423339
Representatividade do faturamento no final de semana	21.59%
Total de vendas nos finais de semana	1833
Total de vendas nos dias de semana	5460
Total de vendas no período	7293,00
Representatividade das vendas no final de semana	25.13%
Produto mais comprado online (ID)	626664333563363
Dia com maior arrecadação	15/8/2018 - Quarta

Tabela 3

4.3 PÁGINAS VISUALIZADAS E COMBINAÇÃO DE DADOS

O conjunto de dados das páginas visualizadas contém um total de mais de 3 milhões de linhas e as 7 colunas mostradas na seção **2.1**. Como é um conjunto de tamanho muito grande tivemos que usar uma abordagem alternativa para manipulação dos dados. Utilizamos a API PySpark para acessar o ambiente Spark e lidar com o grande volume de dados.

Notamos um grande número de dados faltantes nas seguintes colunas do conjunto de dados:

- category_id -1014834 nulos
- customer_id – 3371775 nulos
- on_product_id – 2047370 nulos

Sendo o com maior número de dados ausentes o customer_id.

Todo visitante do site tem uma identificação, adotamos a hipótese de que um cliente, identificado ou não, pode acessar o site diversas vezes e com identificação de visitante diferente.

Com base na questão 4 mostrada na seção 3 **(é comum escolher online e terminar a compra na loja física?)** geramos uma lista de IDs de visitantes (visitor_id) únicos e filtramos esses valores com base nos dados de compras para gerar uma lista de visitantes que não realizaram compra online.

O número total de visitas que não emitiu uma ordem de compra é **858839** e desse número apenas **3014** clientes são identificados. Utilizamos a identificação desses clientes (customer_id) para verificar quantos destes estiveram em alguma loja física no período, e o resultado foi de **257** clientes. É

um valor baixo comparado à quantidade de visitas realizadas no site, porém o baixo número de identificação não nos garante uma estimativa mais precisa uma vez que o número de IDs únicos diferentes de nulo dos clientes nas lojas físicas é de **9902** e o número de clientes que acessaram páginas identificados é aproximadamente **3.2x** menor que esse valor.

O número de visitas às páginas por dispositivo é:

- Desktop – 1353749
- Mobile - 2098791

A quantidade de usuários identificados por dispositivo é expressa no gráfico abaixo:

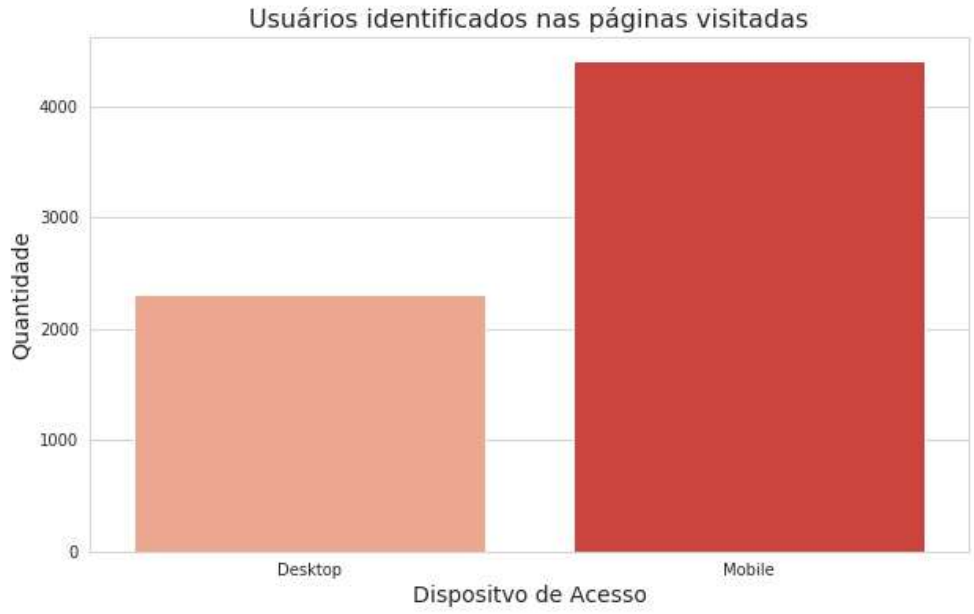

Figura 9

Notamos que a quantidade de usuários identificados nas páginas visitadas é maior quando acessados por mobile, porém o percentual relativo ao total de acessos da categoria é um número muito baixo e muito próximo à outra categoria.

- Percentual de usuários registrados Desktop: **0.17%**
- Percentual de usuários registrados Mobile: **0.20%**

O fragmento de tabela abaixo mostra alguns dados obtidos nessa primeira análise.

Número de clientes identificados que compraram em ambos serviços	457
Quantidade de visitantes únicos na Loja Virtual	864938
Quantidade de visitantes que não realizaram compras	858839
Quantidade de visitantes não compradores identificados	3014

Tabela 4

Para realização da questão 5 adotamos uma abordagem puramente analítica uma vez que a questão solicitava uma campanha direcionada para um cliente específico e identificado pois havia a necessidade de envio de um cupom para o mesmo que só se tornaria possível na posse da identificação.

A hipótese adotada é que um carrinho está abandonado quando, para um ID de visitante que **não** realizou compra online existe um ID de produto diferente de nulo relativo a este visitante.

Primeiramente filtramos os visitantes que não haviam escolhido nenhum produto e excluímos da nossa análise. Após isso geramos uma lista de visitantes que selecionaram algum produto e não compraram nada na loja virtual. A quantidade de visitas com algum produto é de **462521**. A figura abaixo mostra a quantidade de visitas com abandono de carrinho por dispositivo.

Figura 10

Desse número filtramos quantas pessoas tinham identificação, ou seja, estavam registradas/logadas na sua conta na loja virtual. O número de clientes identificados que abandonaram o site com algum produto no carrinho foi de **2566**. A partir desse número podemos levantar algumas possíveis situações:

- Abandono de carrinho ocorre devido à uma UX (User Experience) ruim na página de cadastro/login.
- Usuários já registrados e conectados quando voltam a acessar o site após algum tempo perdem a sessão (baixo tempo em cache).
- Caso seja possível realizar o cálculo do preço do frete antes de se conectar, o usuário o realiza e observa valores altos e desiste da compra.

Complementar aos dados coletados descritos acima, geramos a seguinte tabela.

Dados Gerais	Valores
Faturamento Total no período (R$)	19389777
Número Total de vendas no período	21912,00
Número de clientes identificados que compraram em ambos serviços	457
Quantidade de visitantes únicos na Loja Virtual	864938
Quantidade de visitantes que não realizaram compras	858839
Quantidade de visitantes não compradores identificados	3014
Produto mais visualizado na Loja Virtual	626334343231306
Quantidade de visitantes com carrinho com algum produto abandonado	462521
Quantidade de clientes identificados com carrinho com produto abandonado	2566

5.0 CONCLUSÃO

A abordagem consistiu basicamente da aplicação de métodos simples de limpeza e manipulação de dados, com esses foi possível extrair informações relevantes sobre os conjuntos de dados.

5.1 RESULTADOS IMPORTANTES E RESPOSTAS

Como mostrado na seção 3 foi solicitado explicitamente algumas respostas a serem extraídas do conjunto de dados.

1) O faturamento total no período foi de R$ 19389777,00. Esse valor representa o faturamento da loja virtual somado ao faturamento das lojas físicas.

2) O produto mais comprado online, em quantidade, foi o com o número de identificação 626664333563363.

3) O percentual de gastos dos cariocas nos finais de semana representa 17.76% do faturamento total das lojas físicas no estado e o percentual de número de compras representa 20.15% do número total de compras nas lojas físicas do estado. No Rio de Janeiro o faturamento maior ocorreu durante os dias da semana no mês de agosto de 2018, principalmente sextas e quintas feiras obtiveram maior média de faturamento.

4) Poucas pessoas visualizaram o site, não compraram online e realizaram a compra na loja física. Além disso poucas pessoas realizaram compras nos dois ambientes.

5) A partir dos dados coletados, a melhor opção para o time de marketing é elaborar uma campanha que incentive a compra online de pessoas que abandonaram o carrinho, pois, além de haver poucas pessoas identificadas para enviar o cupom, existem poucos clientes que utilizam ambos serviços. Um cliente que escolhe por uma compra online, deseja realiza-la deste modo e fazer com

que ele vá na loja física pode gerar uma experiência ruim fazendo com que não volte a comprar no futuro.

Além disso, a necessidade de fidelização dos clientes não identificados é um passo importante uma vez que esses dados são cruciais para a empresa melhorar o relacionamento com o cliente com ações como:

- Promoções.
- Divulgação de novos produtos.
- Serviços.
- Brindes.

A partir desta proposta podemos adotar medidas como:

- Revisão na política de cadastro.
- Diferentes maneiras de Login que facilitem acesso (Redes Sociais, etc).
- Revisão na política de frete.
- Benefícios para clientes cadastrados.

Todo material utilizado referente as análises presentes neste relatório, como notebooks, tabelas, imagens e arquivos auxiliares podem ser encontrados no GitHub e em breve no site pessoal do autor.

6.0 REFERÊNCIAS

https://pandas.pydata.org/

https://numpy.org/

https://matplotlib.org/

https://seaborn.pydata.org/

https://stackoverflow.com/

https://spark.apache.org/docs/latest/api/python/index.html