MODELO PARA A ENTREGA DAS ATIVIDADES

COMPONENTE CURRICULAR:	PROJETO APLICADO I
NOME COMPLETO DO ALUNO:	Clayton dos Santos Lira Lorena Vaz Cord Tiago Clemente Rodrigues Lucas Vaz de Castro Oliveira
RA:	10416054 10424700 10423746 10424623

Atenção: Toda atividade deverá ser feita com fonte Arial, tamanho 11, espaço de 1,5 entre as linhas e alinhamento justificado entre as margens.

UNIVERSIDADE PRESBITERIANA MACKENZIE

Tecnologia em ciência de dados

Estudo do cenário de e-commerce no Brasil

São Paulo

2024

Sumário

Introdução

O comércio eletrônico, ou e-commerce, tem se estabelecido como uma das formas mais relevantes e dinâmicas de comércio em todo o mundo, impulsionando a economia digital e transformando os padrões de consumo. No Brasil, esse setor tem experimentado um crescimento significativo, influenciado por diversos fatores, como o aumento da conectividade, a expansão do acesso à internet e mudanças nos hábitos de consumo da população.

Esta análise se propõe a investigar o e-commerce no Brasil no período de 2016 a 2018, utilizando dados disponibilizados pela plataforma Olist, empresa esta que desempenha um papel central no ecossistema do e-commerce brasileiro.

Diante desse contexto, o problema de pesquisa deste estudo é investigar os padrões de compra dos consumidores no cenário do comércio eletrônico brasileiro, utilizando os dados fornecidos pela Olist no período de 2016 a 2018. A análise desses padrões e comportamentos de compra possibilitará uma compreensão mais profunda das tendências de mercado, fornecendo insights valiosos sobre o desenvolvimento do e-commerce no Brasil.

Cronograma de trabalho

Atividades	Março				Abril				Maio			
	1	2	3	4	1	2	3	4	1	2	3	4
Pesquisa bibliográfica												
Delimitação do tema												
Busca da base de dados												
Criação do repositório												
Elaboração dos objetivos												
Elaboração e refinamento da proposta analítica												
Codificação dos scripts da análise												
Elaboração e refinamento de elementos gráficos												
Elaboração do documento escrito												
Confecção da apresentação de slides												
Gravação e edição da apresentação												
Entrega final												

Objetivo de estudo

O tema escolhido é sobre a situação do e-commerce no Brasil a partir de datasets com diversas informações comerciais disponibilizados pela Olist no site Kaggle. Os datasets contém informações úteis como quais os métodos de pagamentos utilizados, valores médios de fretes e de produtos comprados, quais as regiões do país que mais realizaram compras e muito mais. O objeto de estudo será cenário do e-commerce no Brasil durante as datas de 2016 a 2018 (que são os dados disponibilizados), quais métricas envolvem esse cenários e quais conclusões podemos deduzir a partir dos dados que serão analisados.

Empresa

A Olist é uma empresa que oferece um ecossistema de soluções para negócios. Dentre os serviços ofertados, está a Olist store, um canal de vendas digital que atua em várias lojas e marketplaces disponíveis no Brasil (Amazon, Shopee, Magalu, Americanas, etc). Por intermédio da Olist, outras empresas podem anunciar seus produtos no site de grandes redes varejistas com suporte a gestão, logística e atendimento ao consumidor. Sendo assim, a Olist não possui produtos físicos, mas permite que lojistas e fabricantes se associem ao seu nome para alavancar suas vendas, uma vez que a conta Olist é muito forte nesses canais de vendas. Dessa forma, a Olist ganha uma porcentagem das vendas e o parceiro consegue alavancar seu negócio e aumentar a volumetria mensal de vendas. Seu site oficial pode ser acessado clicando aqui.

Problema de pesquisa

Investigar os padrões de compra dos consumidores no cenário do comércio eletrônico no Brasil, utilizando dados da plataforma Olist no período de 2016 a 2018. A análise busca fornecer insights sobre os comportamentos dos clientes durante esse período, permitindo uma compreensão mais profunda das tendências de compra dos consumidores.

Referências de aquisição do dataset

O dataset foi disponibilizado pela empresa proprietária dos dados por meio da plataforma Kaggle, podendo ser consultado através do endereço https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. O Kaggle é uma conhecida comunidade online e plataforma de competições voltada para profissionais e entusiastas da ciência de dados. O referido dataset foi disponibilizado sob licença pública creative commons do tipo CC BY-NC-SA 4.0 (Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional). Dessa forma, é permitido reproduzir e compartilhar tais dados originais ou adaptados, no todo ou em parte, somente para fins não comerciais.

Descrição do dataset e metadados

Segundo informações do próprio proprietário dos dados, os conjuntos de dados disponibilizados possuem informações acerca de clientes, sua localização e de vendedores, valores de frete e produtos, quais categorias de produtos mais compradas, métodos de pagamentos utilizados e muito mais. Os dados possuem informações de cerca de 100.000 pedidos feitos entre 2016 e 2018 em vários marketplaces no Brasil. Os dados disponibilizados são reais, mas mantém anonimato de clientes e parceiros comerciais, os quais tiveram seus nomes substituídos por nomes de grandes casas da série de televisão Game of Thrones.

Os dados são compostos por 9 datasets no formato csv, sendo estes nomeadamente:

1. olist_customers_dataset.csv

Descritivo: Contém informações referentes aos compradores e sua localização.

Linhas: 99.441

Atributos:

- customer_id (string)
- customer_unique_id (string)
- customer_zip_code_prefix (integer)
- customer_city (string)
- customer_state (string)

2. olist_geolocation_dataset.csv

Descritivo: Contém informações sobre códigos postais brasileiros, assim como latitudes e longitudes correspondentes.

Linhas: 1.000.163

Atributos:

- geolocation_zip_code_prefix (integer)
- geolocation_lat (float)
- geolocation_lng (float)

- geolocation_city (string)
- geolocation_state (string)

3. olist_order_items_dataset.csv

Descritivo: Contém dados dos itens comprados dentro de cada pedido.

Linhas: 112.651

Atributos:

- order_id (string)
- order_item_id (integer)
- product_id (string)
- seller_id (string)
- shipping_limit_date (datetime)
- price (float)
- freight_value (float)

4. olist_order_payments_dataset.csv

Descritivo: Possui informações sobre as opções de pagamento de cada pedido.

Linhas: 103.886

Atributos:

- order_id (string)
- payment_sequential (integer)
- payment_type (string)
- payment_installments (integer)
- payment_value (float)

5. olist_order_reviews_dataset.csv

Descritivo: Contém avaliações feitas pelos compradores.

Linhas: 104.719

Atributos:

- review_id (string)
- order_id (string)
- review_score (integer)
- review_comment_title (string)
- review_comment_message (string)
- review_creation_date (datetime)
- review_answer_timestamp (datetime)

6. olist_orders_dataset.csv

Descritivo: Contém todas as informações a respeito dos pedidos.

Linhas: 99.441

Atributos:

- order_id (string)
- customer_id (string)
- order_status (string)
- order_purchase_timestamp (datetime)
- order_approved_at (datetime)
- order_delivered_carrier_date (datetime)
- order_delivered_customer_date (datetime)
- order_estimated_delivery_date (datetime)

7. olist_products_dataset.csv

Descritivo: Inclui dados sobre os produtos vendidos através da Olist.

Linhas: 32.951

Atributos:

- product_id (string)
- product_category_name (string)

- product_name_lenght (integer)
- product_description_lenght (integer)
- product_photos_qty (integer)
- product_weight_g (integer)
- product_length_cm (integer)
- product_height_cm (integer)
- product_width_cm (integer)

8. olist_sellers_dataset.csv

Descritivo: Contém dados sobre os vendedores.

Linhas: 3.095

Atributos:

- seller_id (string)
- seller_zip_code_prefix (integer)
- seller_city (string)
- seller_state (string)

9. product_category_name_transl

Descritivo: Contém dados usados apenas para traduzir a categoria do produto de português para inglês.

Linhas: 71

Atributos:

- product_category_name (string)
- product_category_name_english (string)

Análise exploratória

Uma das partes mais importantes deste trabalho, senão a mais importante, é o estudo dos dados e quais insights podemos tirar a partir deles. Para que tal ocorra é importante resgatar e utilizar bastantes conceitos que vimos e estamos acompanhando em aula como estudo de programação e algoritmos com python e estatística.

Nosso estudo se dará a partir das bases disponibilizadas e de seus conteúdos. Através da estatística e de python, realizaremos o estudo usando do Google Colab que será upada em nosso Github. Todo o estudo e explicação estará nos próprios códigos e notas do Colab.

Para a primeira parte deste trabalho, realizamos o tratamento dos dados e das bases (os códigos e detalhamentos podem ser encontrados no Github). Para um estudo estatístico ser realizado sem enviesamento, é importante tratar os dados adequadamente. Para tal, selecionamos as bases que utilizaremos nos estudos e as tratamos para eliminar ruídos. Nesse tratamos, realizamos as seguintes tarefas de tratamento:

- Verificamos as colunas e linhas dos dataframes com os comandos head e shape;
- Renomeamos as colunas que serão utilizadas em fórmulas e gráficos;
- Tratamos a tipagens dos dados para funcionar corretamente em fórmulas e códigos, ajustando as tipagem de datas;
- Por fim, também tratamos valores missing/NaN, pois podem enviesar nossos estudos. Como a taxa de valores NaN não passaram de 3% do total em algumas bases, as substituímos por sua respectiva moda.

Uma vez que os dados estejam tratados, partiremos para a análise exploratória de fato para verificar quais insights podemos tirar a partir das mesmas. Buscaremos algumas respostas por meio de algumas perguntas como:

1. Quais os meios de pagamentos mais utilizados pelos consumidores?
2. Quais categorias de produtos mais comprados?
3. Qual a distribuição dos preços dos produtos e fretes?
4. Em quais regiões/cidades do país os vendedores mais se concentraram?

Entre outras perguntas. Utilizaremos a estatística e plotagem de gráficos para elucidar as informações.

A partir dos dados e bases tratadas, começamos a analisar algumas informações disponíveis em cada base do nosso conjunto de dados.

Base orders_items

Começando pela base “orders_items”, que possuí informações importantes para nós como preços dos pedidos (price) e valores de frete (freight_value).

A partir de algumas análises realizadas na base de pedidos, encontramos algumas conclusões (indicamos consultar a documentação de códigos disponível em nosso github para maior detalhamento):

- Entre setembro/2016 a abril/2020:

1. É possível verificar que os produtos possuem um preço médio de R$ 120,65;
2. Um valor médio de frete de R$ 19,99 sendo a maioria entre R$ 10 e 29,99.

- Anual:

1. O valor médio do frete não variou muito, ficando perto da casa dos R$20;
2. Houve queda, entretanto, pequena, na média dos preços dos pedidos entre jan/2017 até jan/2019.

Base orders_reviews

Esta base, que diz respeito a comentários e análise dos pedidos de usuários, também traz algumas informações interessantes de se analisar, como nota deixadas para seus pedidos e comentários realizados.

Assim como anteriormente, recomenda-se consultar o github para maiores detalhamentos e acesso aos códigos. Analisando alguns dados sobre as notas dos usuários, pôde-se verificar que:

1. 57,8% dos usuários deram nota 5 em seus comentários, seguido de 19,3% de nota 4. Entretanto, mais de 10% dos compradores deram nota 1 (11,5%);
2. Isso significa que a base de clientes é composta por 77,1% de compradores promotores (que deram notas 5 e 4) e 14,7% de detratores (que deram notas 1 e 2).

A respeito dos comentários, por se tratarem de muitos e com longas frases, a forma mais prática e rápida de analisar é mediante uma nuvem de palavras, em que pôde-se perceber que:

1. As palavras "produto", "prazo" e "entrega" foram as palavras mais usadas pelos usuários. Ainda, palavras como "qualidade" e "muito bom" indicam que a maioria dos usuários ficaram satisfeitos com os serviços oferecidos (produto comprado e entrega). Essa alegação pode ser confirmada não somente pela nuvem de palavras, mas também pela quantidade de promotores, como visto anteriormente. (A data dos comentários compreende-se entre out/2016 a ago/2018).

Por fim, concluí-se que durante o período em que ocorreram os comentários, a Olist conseguiu gerar uma boa impressão aos clientes, garantindo bons prazos de entregas (segundo a maioria e a nuvem de palavras) e um bom nível geral de serviço, uma vez que a grande maioria dos usuários foram promotores da marca (77,1%).

Base orders_payments

Esta base traz informações a respeito de quais formas de pagamentos foram utilizadas pelos clientes, número de parcelas assim como o valor de cada uma. Entretanto, o dado que mais nos interessa é quais meios de pagamentos foram utilizados.

Os meios de pagamentos mais utilizados foram:

- Cartão de crédito, com quase 74% e
- Boleto, com 19%.

Como os dados são de antes de 2021, ano em que o pix foi implementado, é natural que o cartão de crédito fosse o meio de pagamento mais utilizado online. Atualmente (2024), acredito que a taxa de uso do cartão de crédito ainda seja a maioria, mas vem perdendo espaço para o pix.

Base orders

Essa base traz informações acerca dos pedidos, como situação do pedido, data de entrega e data estimada de entrega, por exemplo. Olhando para as datas de entregas, vimos que se compreenderam entre out/2016 a out/2018.

No geral, pedido efetuado foi pedido entregue! Naturalmente, uma empresa comprometida com os clientes, entrega os produtos comprados. Com a Olist não foi diferente, no período analisado, 97% das entregas foram efetuadas. Assim como na estatística, é impossível controlar 100% das entregas e houve variabilidades. Entre as ocorridas, destacaram-se "shipped"/despachado (1,1%) e cancelados (0,7%).

Base products

Esta base, talvez a mais simples de todas, traz somente informação da categoria de produtos mais pedidos. Entre as muitas categorias encontradas, 73 mais especificamente, selecionamos as 10 mais comuns entre os clientes e podemos verificar que, no top 3 mais pedidos estão:

- cama_mesa_banho, com 3639 pedidos;
- esporte_lazer, com 2867 pedidos;
- moveis_decoracao, com 2657 pedidos.

Esses valores correspondem, respectivamente, a 11%, 8,7% e 8,6% do total de pedidos realizados (32951).

Base sellers

Esta base traz informações dos vendedores, como de quais estados e cidades são. A maioria dos vendedores que atuaram pela Olist pertencem aos estados de:

- São Paulo (59,7%),
- Paraná (11,3%)
- e Minas Gerais (7,9%).

Portanto, mais da metade dos vendedores atuaram em SP, sendo as regiões sudeste e sul as mais presentes.

Base customers

Esta base traz informações dos clientes, como de quais estados e cidades são. Os clientes apresentaram uma divisão regional bem definida:

Em primeiro lugar, a região sudeste apresentou os maiores valores:

- São Paulo (42%),
- Rio de Janeiro (12,9%)
- e Minas Gerias (11,7%.

Com um totalizando 66,7% dos clientes. Em segundo lugar, a região sul apresentou:

- Rio Grande do Sul (5,5%),
- Paraná (5,1%)
- e Santa Catarina (3,6%), totalizando 14,2%.

Conclusão final

Durante todo o período analisado, os preços médios dos pedidos não variou muito ao longo dos anos, ficando entre R$ 120 e 135,00, sendo 75% do total dos pedidos abaixo desse último valor mencionado. Como o valor médio dos pedidos foi baixo, o valor médio do frete também foi, ficando em torno de RS 20,00.

Por se tratar de e-commerce, naturalmente, a maioria dos pedidos foram efetuados por meio de cartão de crédito (74%), seguido de boleto (19%). Acredito que se a base estivesse atualizada até esse ano (2024), o surgimento de pix como meio de pagamento provavelmente estaria entre os três primeiros lugares, uma vez que, foi grandemente aceito e possui uma grande facilidade e agilidade em relação ao boleto.

No geral, os pedidos geraram uma boa nota, formando uma maioria de clientes promotores da marca (notas 5 e 4), atingindo um percentual de 77,1% e 14,7% de detratores (notas 1 e 2). Na nuvem de palavras, foi possível encontrar entre as palavras mais frequentes, palavras como "produto", "prazo" e "entrega", que foram usadas em comentários de elogios, fazendo sentido, já que 97% dos pedidos foram entregues. Palavras como "recomendo" e "muito bom" indicaram que a maioria dos usuários ficaram satisfeitos com os serviços oferecidos (pedidos realizados e entrega). Ainda sobre entregas, entre 3% de não entregues, "shipped"/despachado (1,1%) obteve a maior porcentagem, seguido de cancelados (0,7%).

Dentre os mais de 100 mil pedidos realizados na base de pedidos, as três principais categorias de produtos foram:

1. cama_mesa_banho, com 11,0%;
2. esporte_lazer, com 8,7%;
3. moveis_decoracao, com 8,6%;

A partir desses dados e atrelado com os baixos valores da maioria dos pedidos, é possível concluir que os clientes normalmente realizam pequenas compra como toalhas para banho, lençóis para cama e utensílios para casa de modo geral.

Acerca dos vendedores e clientes, a maioria está concentrada nas regiões sudeste e sul, respectivamente. Entre os vendedores, quase 60% estão concentrados no estado de São Paulo. Entre os clientes, 42% estão em São Paulo e 12,9% no Rio de Janeiro.

Por fim, (acreditamos que a Olist tenha esse dado, somente não o disponibilizou nessas bases) mas um dado faltante e extremamente interessante que poderia ter, é a segmentação dos dados por marketplace/plataforma de vendas (como Mercado Livre, Amazon, Magalu, Shopee, etc). Assim, seria possível utilizar todo o potencial dos dados disponibilizados para verificar qual plataforma está permitindo o maior número de vendas/pedidos, o maior ticket médio, a maior taxa de elogios/reclamações entre os usuários ou que possuem as taxas de entregas mais eficientes (embora esse dado parte mais da transportadora responsável) ou até mesmo, verificar quais produtos e categorias performam melhor em cada plataforma. Seria possível ter uma visão dinâmica de como anda o negócio em cada plataforma, permitindo estratégias e ações mais específicas e assertivas, alavancando mais ainda a grande performance que a Olist já pôde realizar.

Repositório

Endereço para o repositório: https://github.com/lucasvazcastro/Projeto_Aplicado_I