Análise em Dados do E-commerce da Olist

Desafio UFRJ Analyica

Juan Victor Carballo Blanco

12/03/2023

Sumário

1 Introdução

O trabalho em questão consiste em uma análise do conjunto de dados de e-commerce da Olist disponibilizado pelo kaggle. As informações foram criadas entre os anos de 2016 e 2018 consistindo de 8 arquivos de formato "csv" que possuem os nomes:

- olist_customers_dataset
- olist_geolocation_dataset
- olist_order_items_dataset
- olist_order_payments_dataset
- olist_order_reviews_dataset
- olist_orders_dataset
- olist_orders_dataset
- olist_sellers_dataset

Todos esses datasets estão interligados através de chaves primárias ou estrangeiros de acordo com a imagem abaixo:

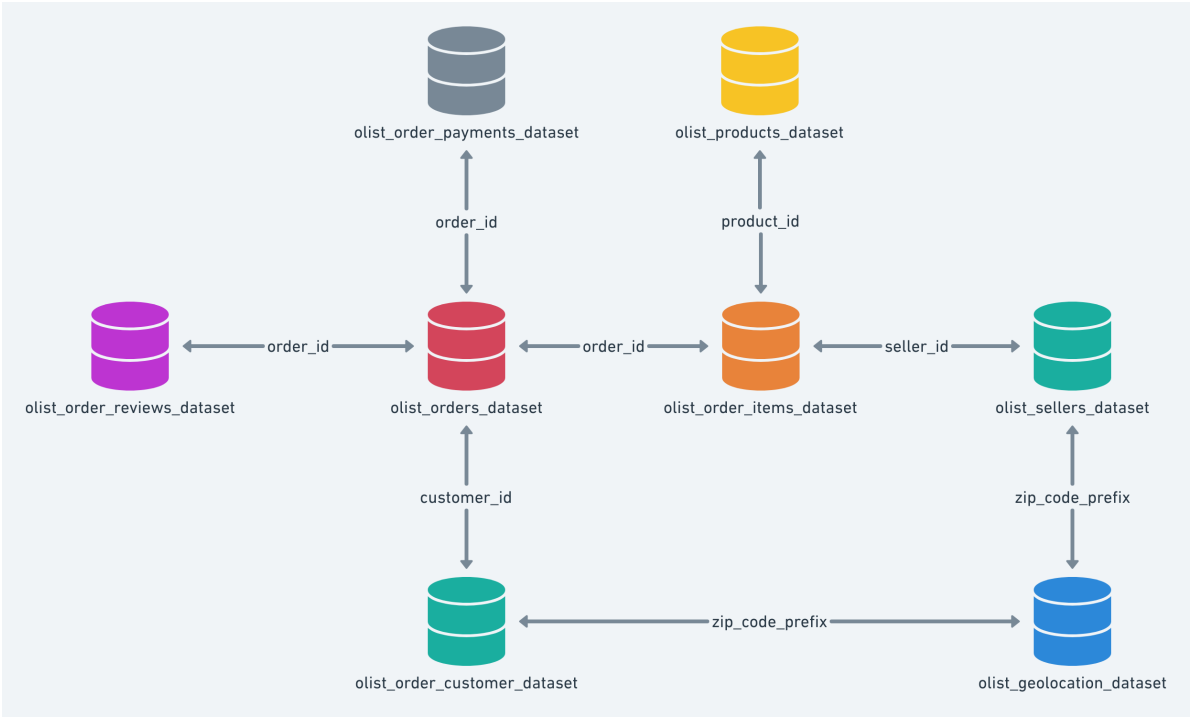


Figure 1: Relacionamento do Conjunto de Dados

O objetivo desse projeto é entender quem são os clientes de Olist e por fim realizar um trabalho de segmentação de clientes.

2 Análise Exploratória de Dados

2.1 Compradores e Vendedores

Analisando a distribuição espacial dos consumidores, podemos verificar que a demanda da Olist se concentra no sudeste, pricipalmente no estado de São Paulo. Isso provavelmente se deve a maior concentração populacional e economica nessa região que, desde a vinda da Familia Real em 1808, foi se consolidando como a região com maior progresso socio-economico do país. O mapa abaixo mostra essa concentração.

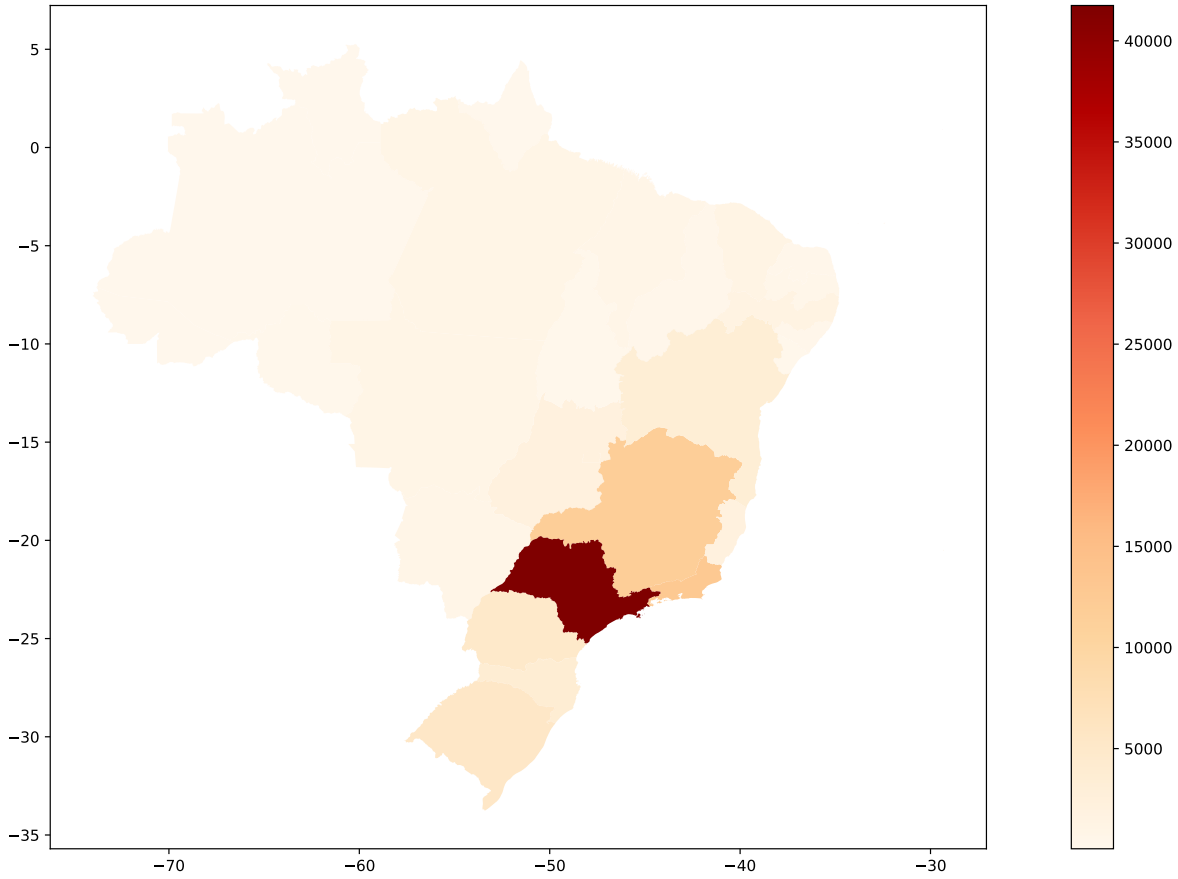


Figure 2: Distribuição de Clientes ao longo do Brasil

Quando analisamos o aspecto da oferta, percebemos que as desigualdades regionais se amplificam, principalmente em São Paulo. Nesse aspecto, podemos ressaltar aspectos de escala principalmente no setor indústrial e logístico. No livro Economia Internacional de Paul Krugman no capítulo 7 é destacado como a produção tende a se concentrar em polos devido a ganhos relacionados a escala. A intuição por trás disso é que custos fixos industriais e logisticos, por exemplo, serão o mesmo valor independentemente se instalado em áreas grande dinamismo economico ou de baixo. Dessa forma, é muito mais eficiente e menos custoso a produção ser concentrado em um só lugar do que descentralizada em vários lugares diferentes. No caso em questão, uma empresa de e-commerce como a Olist consegue reduzir seus custos concentrando a produção em um lugar, já que se beneficiaria desse aumento de escala tanto vinda de si própria como de empresas na região.

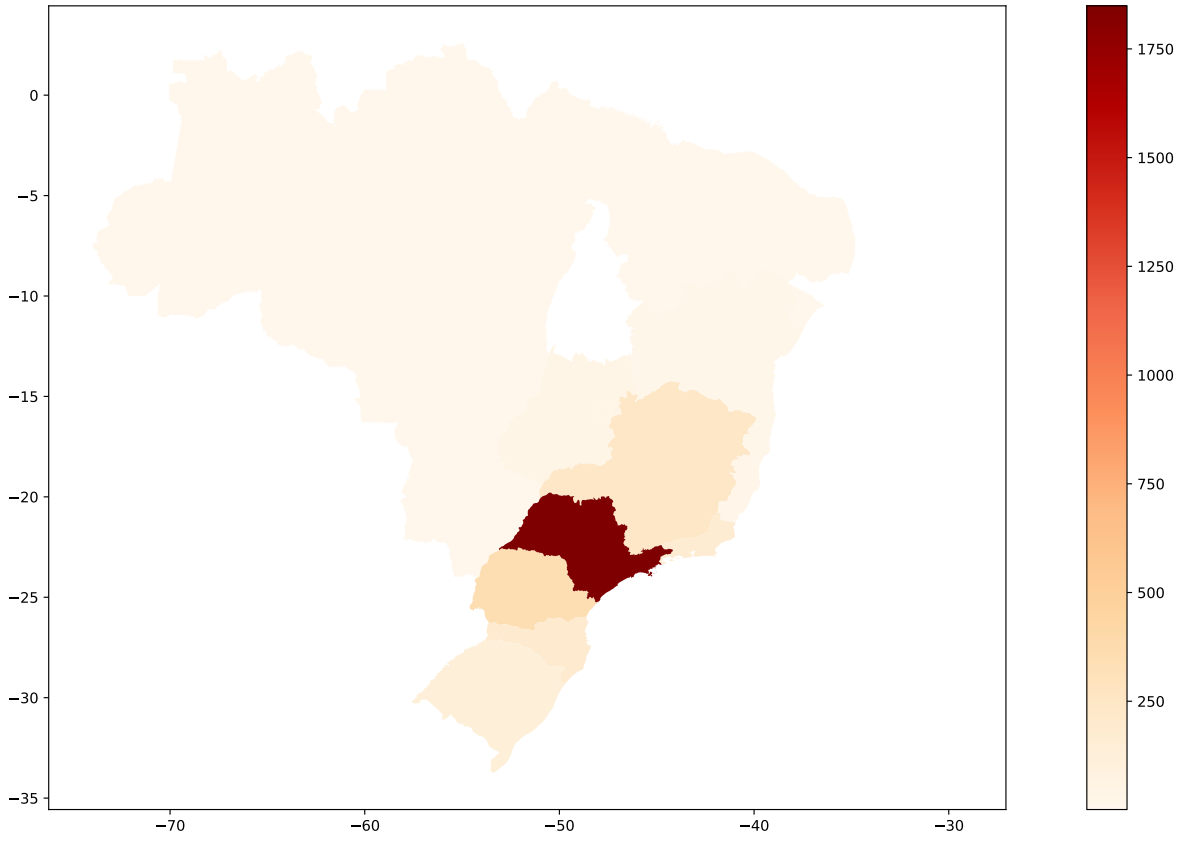
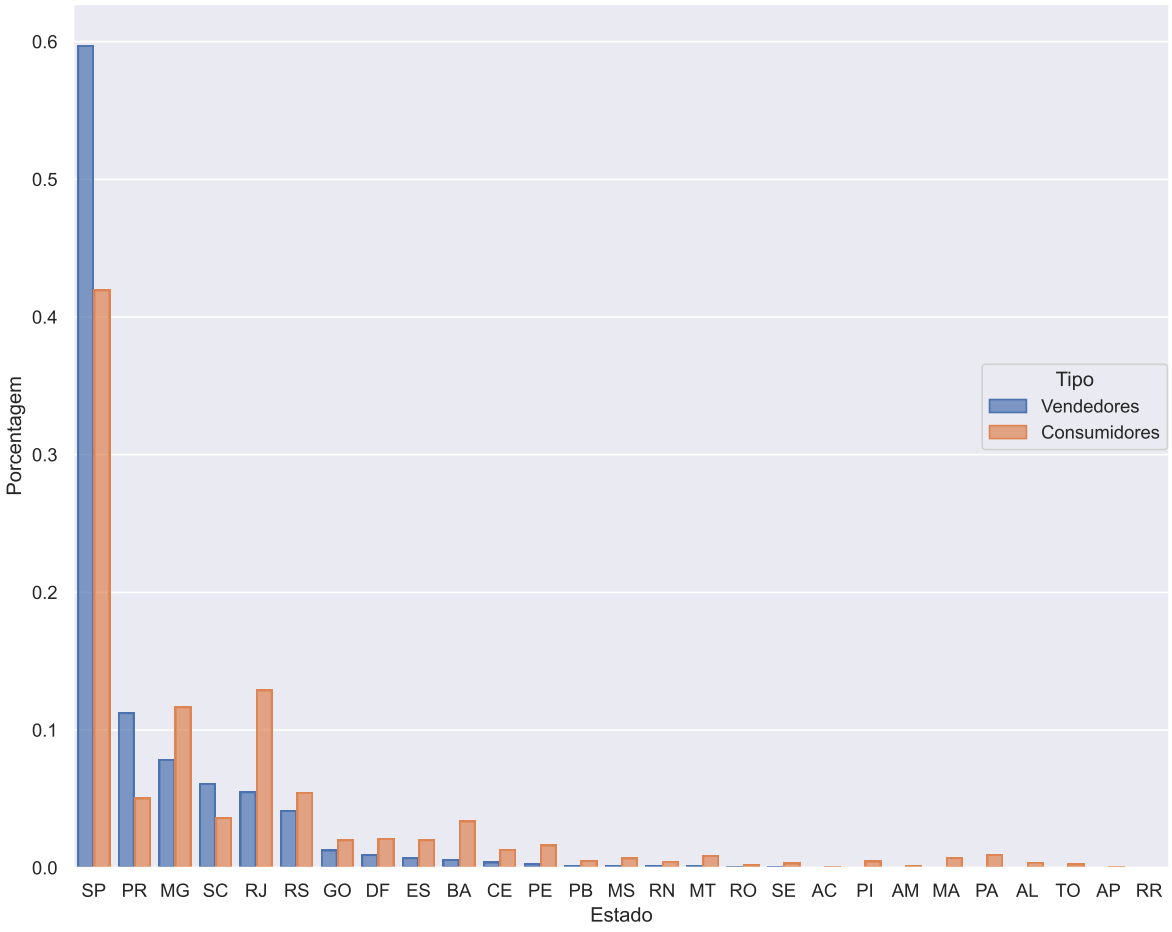


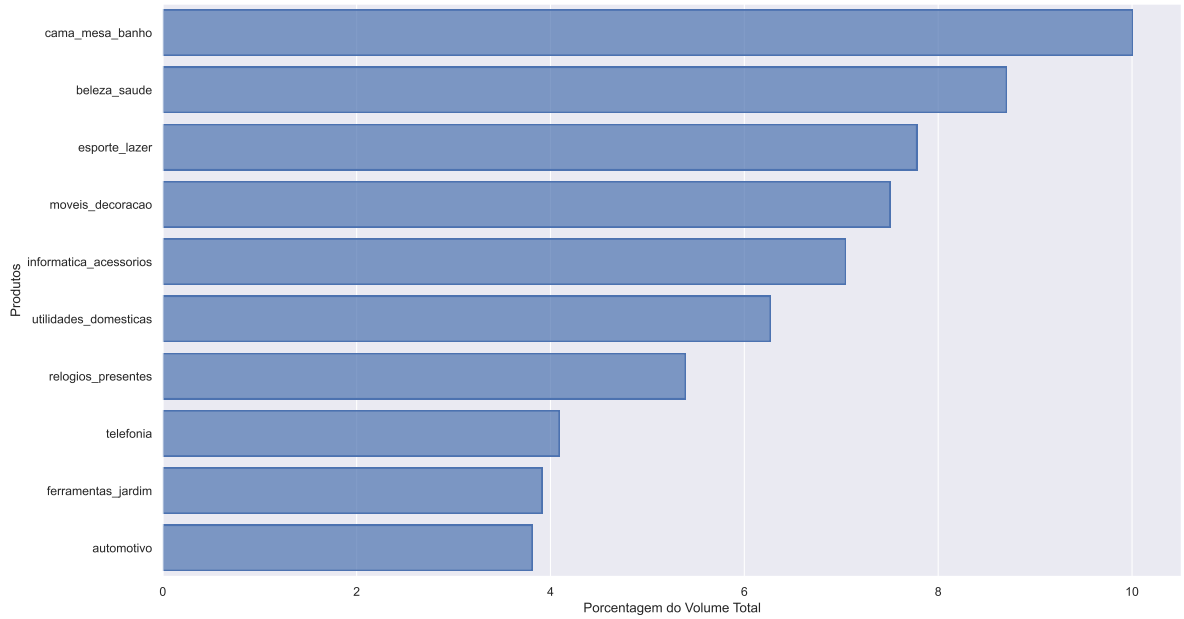
Figure 3: Distribuição de Varejistas ao longo do Brasil

O gráfico abaixo nos mostra novamente essa diferença entre consumidores e vendedores onde temos tanto na oferta quanto na demanda a proporção demandada/ofertada em relação ao volume total. Alguns poucos Estados (São Paulo, Paraná e Santa Catarina) possuem mais vendedores que consumidores. Podemos pensar que, nesse caso, esses estados estivessem "exportando" produtos enquanto os demais estivessem "importando".



2.2 Produto

Os produtos mais vendidos na Olist são produtos relacionados a vida de uma jovem familia brasileira. Infelizmente, não temos acesso a dados demográficos, porém, em vista dos 10 produtos mais vendidos, eu diria que o cliente médio são adultos de mais de 30 anos, pois grande parte desses produtos são relacionados a itens de casa.



Ao analisar a nota dada por cada produto mais vendido, parece que não há nenhuma preferencia dos consumidores. A tabela abaixo mostra essa a nota média por cada produto

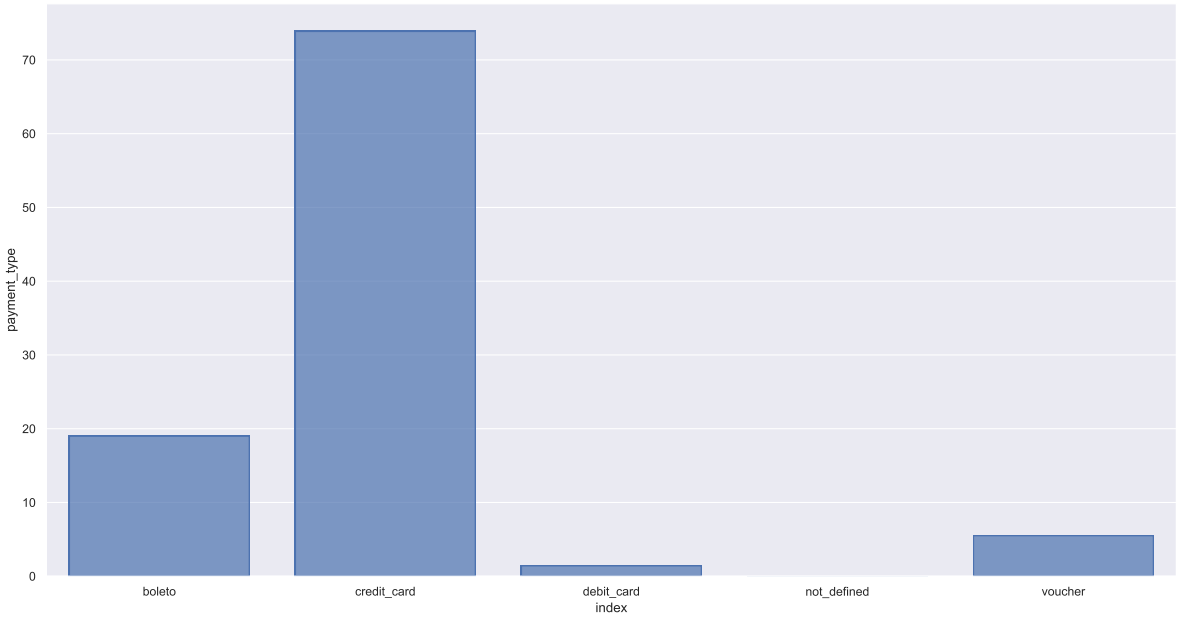
	Produto	Avaliação
0	automotivo	4.07
1	beleza saude	4.14
2	cama mesa banho	3.90
3	esporte lazer	4.11
4	ferramentas jardim	4.04
5	informatica acessorios	3.93
6	moveis decoracao	3.90
7	relogios presentes	4.02
8	telefonia	3.95
9	utilidades domesticas	4.06

Como podemos observar na tabela abaixo, há uma quantidade considerável de outliers no conjunto de dados em relação ao valor total pago. Uma pequena quantidade de consumidores gastaram alguns milhares de reais, enquanto a vasta maioria tem seu ticket médio por volta de 90 reais.

	Media	Desvio Padrão	Min	25%	Mediana	75%	Max
total_cost	140.644059	190.724394	6.08	55.22	92.32	157.9375	6929.31

2.3 Tipo de Pagamento

O tipo de pagamento mais utilizado em compras é, com uma grande diferença, o cartão de crédito, provavelmente devido a sua habilidade de parcelar suas compras.



Apesar disso, cerca de 50% das vendas foram a vista e das que foram parceladas houve uma clara tendencia a um menor parcelamento possível. Esse fenomeno talvez seja devido ao baixo ticket médio do e-commerce, já que valores mais altos tendem a serem pagos em maiores parcelas. As tabelas abaixo mostram esses fenomenos.

	payment_installments	Perc_total
0	1	50.58
1	2	11.95
2	3	10.07
3	4	6.83
4	10	5.13
5	5	5.04
6	8	4.11
7	6	3.77
8	7	1.57
9	9	0.62

3 Modelagem

Meu objetivo é criar um modelo que crie cluster de clientes com base em suas características. Para tal, utilizarei o algoritmo K-Means para o agrupamento e o método Elbow para determinar o número de clusters. Para isso, eu utilizarei o conceito de RFM (Recency, Frequency, Monetary) para criar as features do modelo.

RFM é um modelo de análise de consumidores que utiliza três variáveis para classificar os clientes de acordo com seu comportamento de compra. A variável Recency (R) mede a o quão recente foi a compra dos clientes, a variável Frequency (F) mede a frequência de compra dos clientes e a variável Monetary (M) mede o valor total gasto pelos clientes. Com essas informações é possível melhorar a tomada de decisão de uma empresa. A tabela abaixo desse site nos mostra como sabendo o perfil do consumidor através dessa metodologia pode ajudar na tomada de decisão empresarial

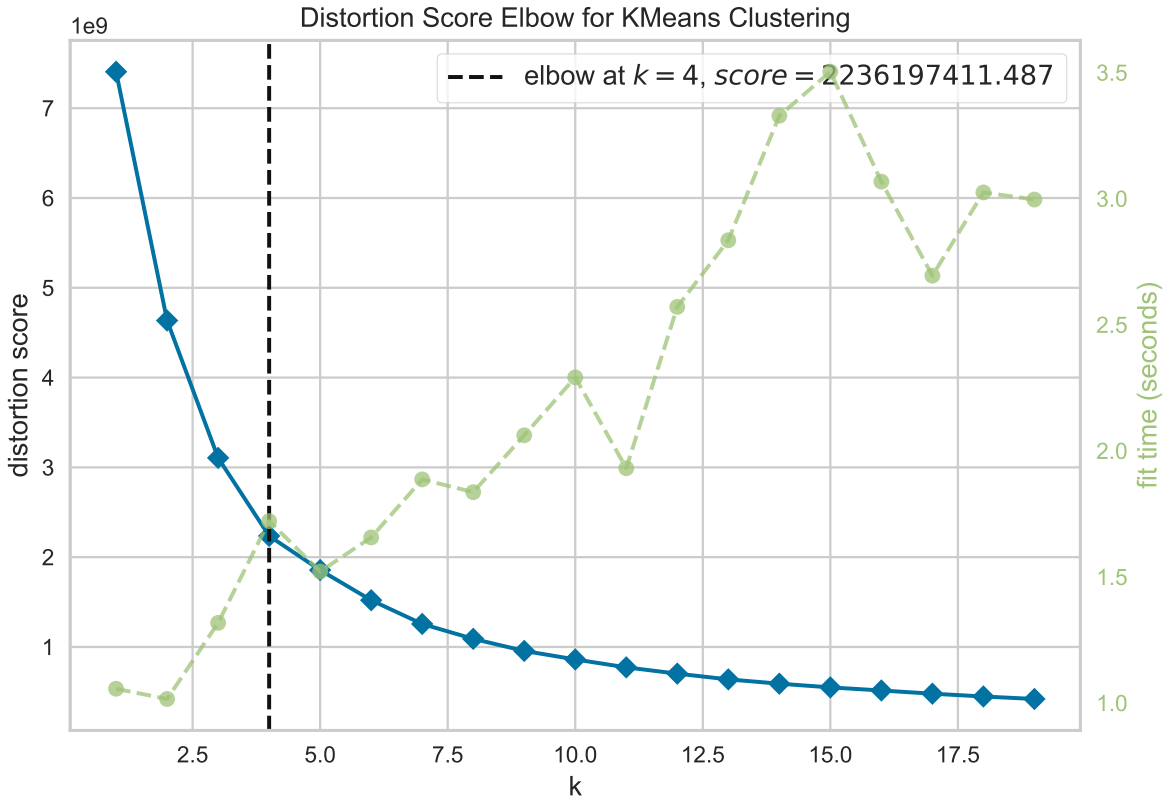
Customer Segment	Activity	Actionable Tip
Champions	Made recent purchases, frequent purchases, high spending	Offer loyalty rewards or exclusive promotions to maintain their loyalty and encourage them to make repeat purchases.
Loyal Customers	Made frequent purchases, but not recently, high spending	Send them personalized offers or promotions based on their past purchases to encourage them to return and make another purchase.
Potential Loyalists	Made recent purchases, but not frequent, high spending	Provide personalized recommendations for related products or services that they may be interested in based on their recent purchase.
Recent Customers	Made recent purchases, but not frequent, low spending	Provide a discount or offer on their next purchase to encourage them to make another purchase and become a potential loyal customer.

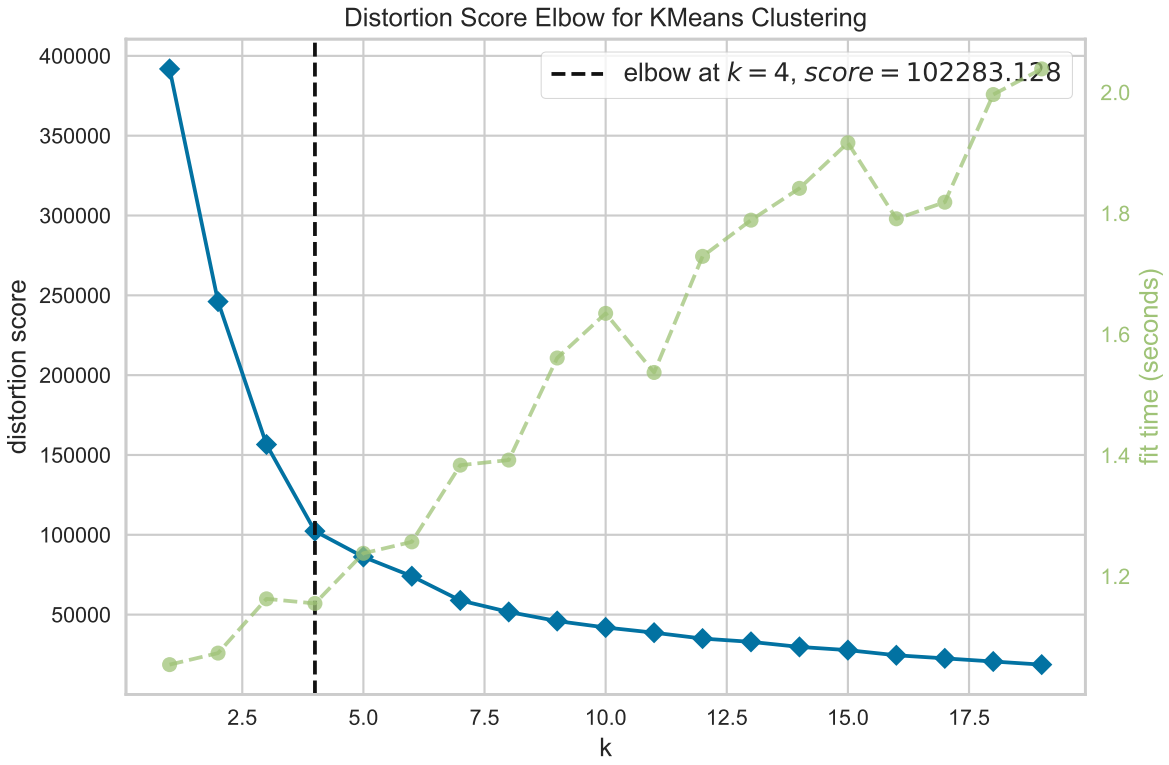
Customer Segment	Activity	Actionable Tip
Promising Customers	Made frequent purchases, but not recently, low spending	Provide them with targeted offers or discounts on products or services they have shown interest in, to encourage them to return and make another purchase.
At-Risk Customers	Made frequent purchases in the past, but not recently, low spending	Send them personalized offers or discounts to encourage them to make another purchase and re-engage with your brand.
Lost Customers	No recent purchases, low frequency, low spending	Re-engage them by sending personalized emails or promotions to encourage them to return and make another purchase.
Lost Cheap Customers	No recent purchases, low frequency, low spending	Offer them a discount or promotion to encourage them to return and make another purchase, but also consider whether it makes sense to focus on acquiring new customers instead.

Geralmente, as variáveis da análise por RFM são posta em uma escala de 1 a 5. Para esse trabalho, foi testado ambos os casos usando o algoritmo K-means. Uma forma de metrificar os clusters é calculando sua distorção que, dentro do contexto do K-means, calcula a diferença entre o centro dos clusters e cada ponto dos dados. Dessa forma, cada vez menor a distorção, mais representativo é o cluster. Evidentemente, a simples minimização da distorção sem limitações na quantidade de clusters nos levaria a uma situação que teriamos a mesma quantidade de observações que de clusters. Nesse contexto, utilizamos o "elbow method" para balancear o trade-off entre minimizar a distorção e minimizar a quantidade de clusters. Essa técnica consiste em calcular a variação da distorção a cada aumento marginal na quantidade de clusters. Quando a variação da distorção for muito pequena, isto é, quando a curva distorção vs quantidade de clusters ser quase plana que parece um cotovelo, seria a quantidade ideal de clusters.

Nos gráficos abaixo foi realizado essa técnica tanto quando os dados de RFM não está em escala de 1 a 5 e quando este está, respectivamente. A conclusão é de que o número ideal de clusters é 4 e que o conjunto de dados ideal para a realização da clusterização é o que teve

sua escala de 1 a 5, de acordo com o valor de distorção. Os gráficos abaixo mostram essa perspectiva.

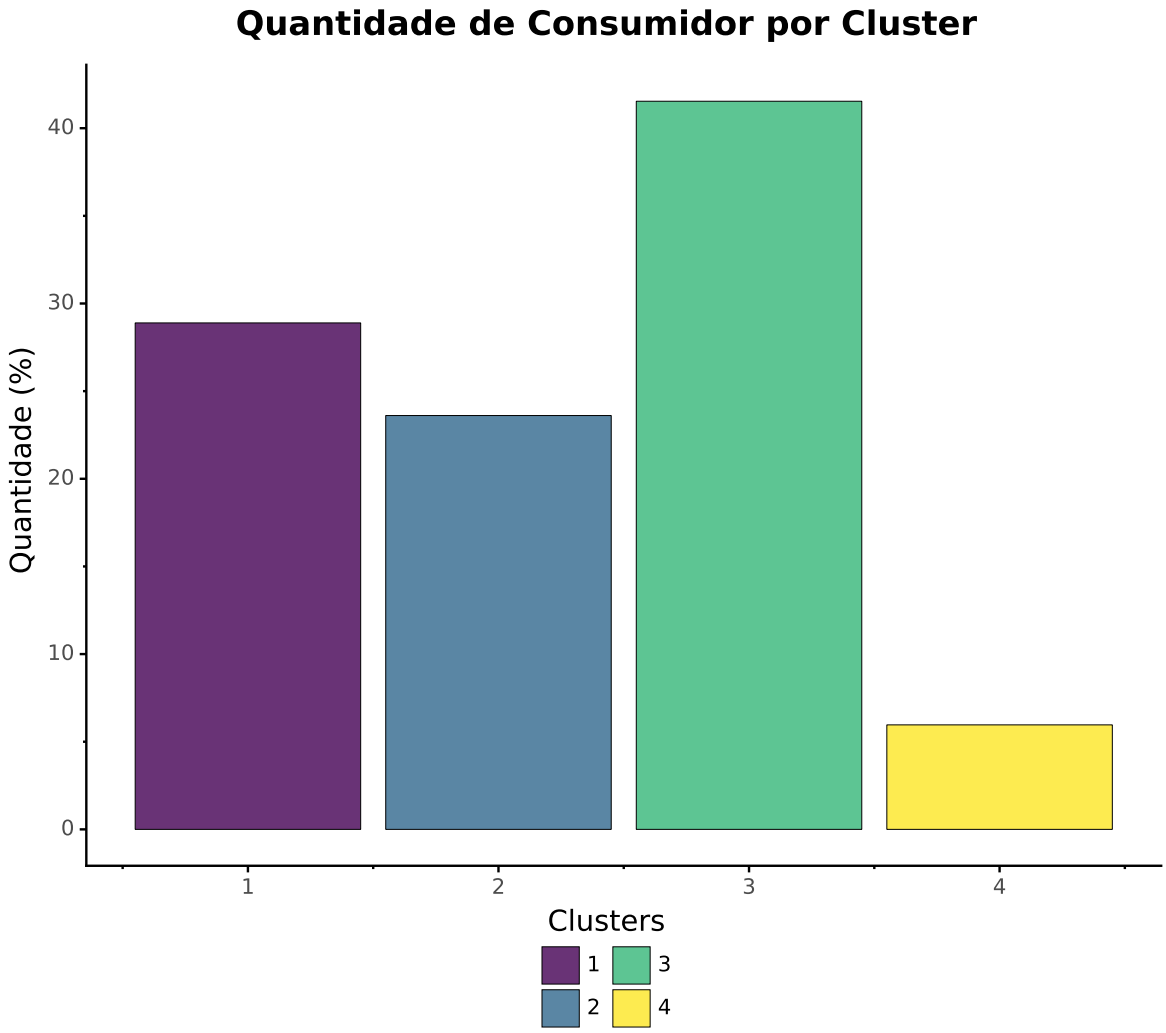




Após verificado quais melhores métricas e conjuntos de dados para realizar o agrupamento dos clientes, podemos classificar os clientes em 4 grupos:

- Grupo 1: Clientes que gastaram muito dinheiro, mas que não compraram com frequência e que não compraram recentemente.
- Grupo 2: Clientes que compraram pouco, não compraram recentemente e não compraram com frequência.
- Grupo 3: Clientes que compraram recentemente, mas que não compraram com frequência e que gastaram uma quantia média de dinheiro.
- Grupo 4: Clientes que compraram com frequência, compraram uma quantidade razoável de dinheiro e que compraram relativamente recentemente.

Atualmente na empresa, como podemos observar no gráfico abaixo, o cluster mais significativo é o 3 e o menos significativo é o 4. Dessa forma, podemos concluir que a Olist conseguiu chamar a atenção de novos clientes que gastam uma boa quantidade de dinheiro, já que a única diferença entre o 3 e o 4 grupo é a frequencia. Dado esse contexto, em vista o paradigma RFM, é evidente que é o ideal para a marca é criar produtos que aumente fidelização de seus clientes como pacotes de assinatura e melhorar sistema de recomendação de produtos.


Quantidade de Consumidor por Cluster
Quantidade (%)
40
30
20
10
0
1
2
3
4
Clusters
1
2
3
4