Análise de E-Commerce

Clusterização e Informações

DataMining

Fabricio Almeida da Silva Nunes
Lorran Richardson de Sa Freire
Ismael Weslley Neves de Brito
Mateus do Nascimento Magalhães da Silva

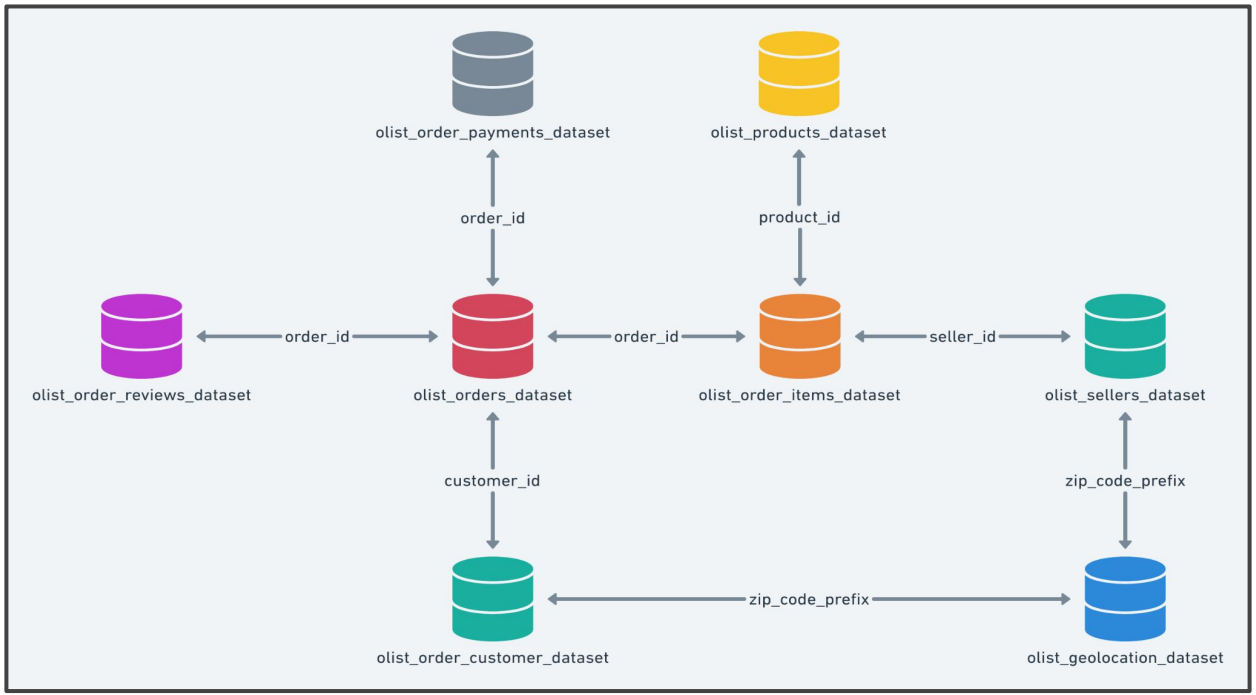
E-Commerce - Olist

Base de dados utilizada:

Para a realização deste trabalho foi utilizado a base de dados disponibilizada por Olist que se encontra em:

Brazilian E-Commerce Public Dataset by Olist

Esquema dos Dados



Bases de dados e seus relacionamentos através de suas chaves primárias

Ferramenta Utilizada

Para as codificações e análises, foi utilizado neste trabalho a linguagem de programação PYTHON através do ambiente de desenvolvimento conhecido como Google COLAB, com bibliotecas específicas para análise de dados como PySpark, Pandas e outras.

O código completo se encontra atualmente disponível em:

https://colab.research.google.com/drive/1mLS9YcrKVaoidK1RfK_PDee_6WwbN2DI?usp=sharing

Um vídeo com a execução completa dos códigos se encontra disponível em:

https://drive.google.com/file/d/1QhohfRA4EJXLhK_Om7OFOTb0l1E9TEkQ/view

Relatório Power Bi

Link para o relatório online: Olist Analysis

Modificações nos Datasets

Para maior precisão nas coordenadas, o dataset Geolocation teve que ser alterado visto que, um mesmo zip_code possui N valores de latitude e longitude diferentes na mesma cidade.

Logo, foi criado um identificador único utilizando o (zip_code + geolocation_state) e os valores de latitude e longitude agora são as médias de todos os valores de mesmo identificador.

Dessa forma pode-se estimar uma latitude e longitude para cada cliente e cada vendedor.

Organização dos dados alterados

order_items : 112650
['order_id', 'order_item_id', 'product_id', 'seller_id', 'shipping_limit_date', 'price', 'freight_value', 'cluster']

order_reviews : 104162
['review_id', 'order_id', 'review_score', 'review_comment_title', 'review_comment_message', 'review_creation_date', 'review_answer_timestamp']

order_payments : 103886
['order_id', 'payment_sequential', 'payment_type', 'payment_installments', 'payment_value', 'cluster']

orders : 99441
['order_id', 'customer_id', 'order_status', 'order_purchase_timestamp', 'order_approved_at', 'order_delivered_carrier_date', 'order_delivered_customer_date', 'order_estimated_delivery_date']

customers : 99163
['cod', 'customer_id', 'customer_unique_id', 'customer_zip_code_prefix', 'customer_city', 'customer_state', 'lat', 'lng', 'cluster']

products : 32951
['product_id', 'product_category_name', 'product_name_lenght', 'product_description_lenght', 'product_photos_qty', 'product_weight_g', 'product_length_cm', 'product_height_cm', 'product_width_cm', 'cluster']

geolocation : 19023
['cod', 'lat', 'lng']

sellers : 3053
['cod', 'seller_id', 'seller_zip_code_prefix', 'seller_city', 'seller_state', 'lat', 'lng', 'cluster']

Dashboard e Análises

Clientes

Análise dos Clientes

Quantidade de clientes

99 Mil

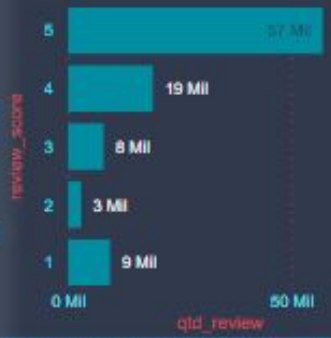
Média da análise de sentimento

6,06

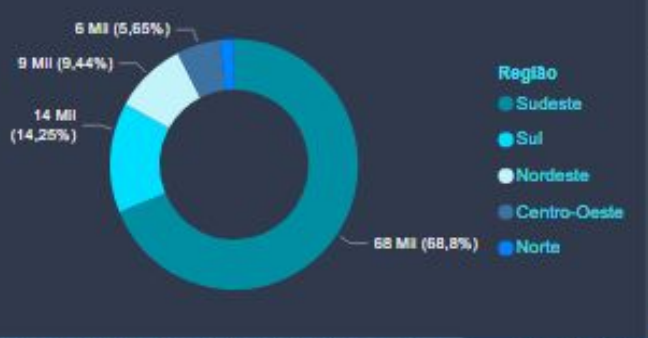
Média do Score

4,09

Quantidade de reviews por nota



Quantidade de cliente por região

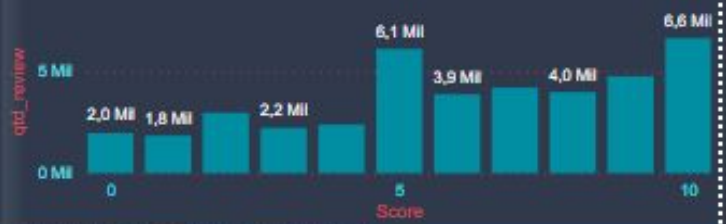


Mapa de calor dos estados por clientes

5 estados com mais clientes, média de preço e frete

5 estados com menos clientes, média de preço e frete

Quantidade de cliente por nota da análise de sentimentos



Linha de tempo da quantidade de clientes

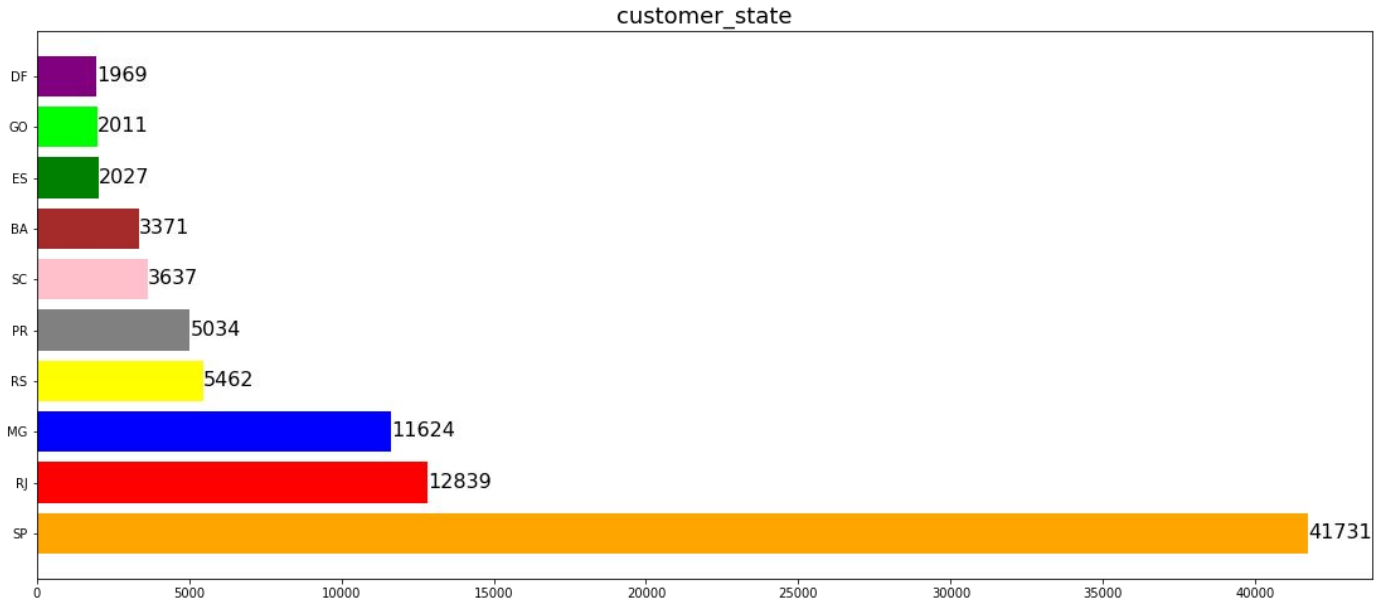
Nuvem de palavras dos comentários das reviews

Clientes com maior e menor valor de compra

Cliente	Estado	Valor de compra	Frete total
1617b1357756262bfa56ab541c47bc16	RJ	R$ 13.440	R$ 224
ec5b2ba62e574342386871631fafd3fc	ES	R$ 7.160	R$ 115
c6e2731c5b391845f6800c97401a43a9	MS	R$ 6.735	R$ 194
f48d464a0baaea338cb25f816991ab1f	ES	R$ 6.729	R$ 193
3fd6777bbce08a352fddd04e4a7cc8f6	SP	R$ 6.499	R$ 228
05455dfa7cd02f13d132aa7a6a9729c6	MG	R$ 5.935	R$ 147
df55c14d1476a9a3467f131269c2477f	RJ	R$ 4.799	R$ 151
24bbf5fd2f2e1b359ee7de94defc4a15	SP	R$ 4.690	R$ 74
e0a2412720e9ea4f26c1ac985f6a7358	GO	R$ 4.600	R$ 210
3d979689f636322c62418b6346b1c6d2	PB	R$ 4.590	R$ 92

Cliente	Estado	Valor de compra	Frete total
9f9d249355f63c5c1216a82b802452c1	RJ	R$ 1	R$ 18
161b6d415e8b3413c6609c70cf405b5a	SP	R$ 1	R$ 18
a790343ca6f3fee08112d678b43aa7c5	SP	R$ 2	R$ 7
184e8e8e48937145eb96c721ef1f0747	SP	R$ 2	R$ 8
77a34d46f6ebd1dcb9b4df9ae5739226	SP	R$ 3	R$ 13
deaf712a6d30217071ce4d2cf4e0ef79	RJ	R$ 3	R$ 15
40ada5e3dc4b3c488f9367c4ba39727a	SP	R$ 3	R$ 17
d2c63ad286e3ca9dd69218008d61ff81	PR	R$ 3	R$ 9
5ac9fcc9259df95cf14d27238b112148	SP	R$ 3	R$ 12
55cd7bfe95dcd698acf176278e14888e	SP	R$ 4	R$ 8

Top 10 estados com mais clientes



A maior concentração de clientes se encontra no estado de São Paulo, com uma quantidade bastante elevada de clientes em relação aos outros estados.

Análise de sentimentos de comentários com Python

	score	txt
0	1	Péssimo
1	1	Não gostei ! Comprei gato por lebre
2	1	Sempre compro pela Internet e a entrega ocorre...
3	1	Nada de chegar o meu pedido.
4	1	recebi somente 1 controle Midea Split ESTILO.
5	1	O produto não chegou no prazo estipulado e cau...
6	1	Produto muito inferior, mal acabado.
7	1	Pedi reembolso e sem resposta até momento
8	1	Este foi o pedido
9	1	comprei tres pacotes de cinco folhas cada de p...

Base de comentários com pontuação de 1 até 5

Preparo de treino/teste com 0.3 de tamanho

```
vect = CountVectorizer(ngram_range=(1, 1))
vect.fit(df.txt)
text_vect = vect.transform(df.txt)

X_train,X_test,y_train,y_test = train_test_split(
text_vect,
df.score,
test_size = 0.3,
random_state = 42
)
```

```
clf = LogisticRegression(random_state=0, solver='newton-cg')
clf = clf.fit(X_train, y_train)

y_prediction = clf.predict(X_test)
f1 = f1_score(y_prediction, y_test, average='weighted')
print("Score: ",f1)



Score:  0.8270471886970595
```

Resultado com 82,7% de precisão
***Precisão pode variar com os testes**

Interface para análise de sentimentos com Python

Predizer sentimento

comentario1: " Odiei esse produto!

comentario2: " "Gostei bastante do conteudo do produto"

comentario3: " produto muito ruim :C

comentario4: " O preco estava otimo

comentario5: " Achei o produto bem péssimo

Mostrar código

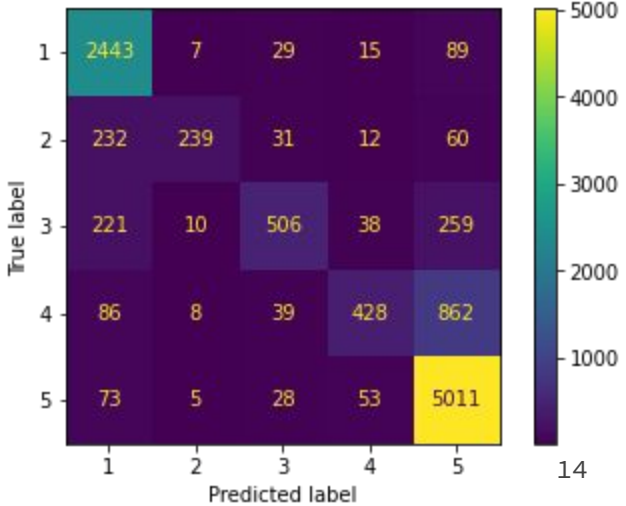
```
( Negativo- ) Odiei esse produto!
( Positivo+ ) "Gostei bastante do conteudo do produto"
( Negativo- ) produto muito ruim :C
( Positivo+ ) O preco estava otimo
( Negativo- ) Achei o produto bem péssimo
```

Basta inserir um comentário e executar para observar o sentimento que será detectado para aquele comentário.

A avaliação pode não ser muito precisa, ainda resultando em avaliações erráticas, mas ainda sim é um ótimo exemplo de como uma análise de sentimentos funciona.

Positivo + -> Nota 5
Positivo -> Nota 4
Neutro -> Nota 3
Negativo -> Nota 2
Negativo- -> Nota 1

Podemos observar que os extremos são mais fáceis de prever o sentimento corretamente, ou seja, é identificar quando a pessoa está muito positiva ou muito negativa



***Matriz de confusão Pode variar**

Quantidade de avaliações separadas por pontuação

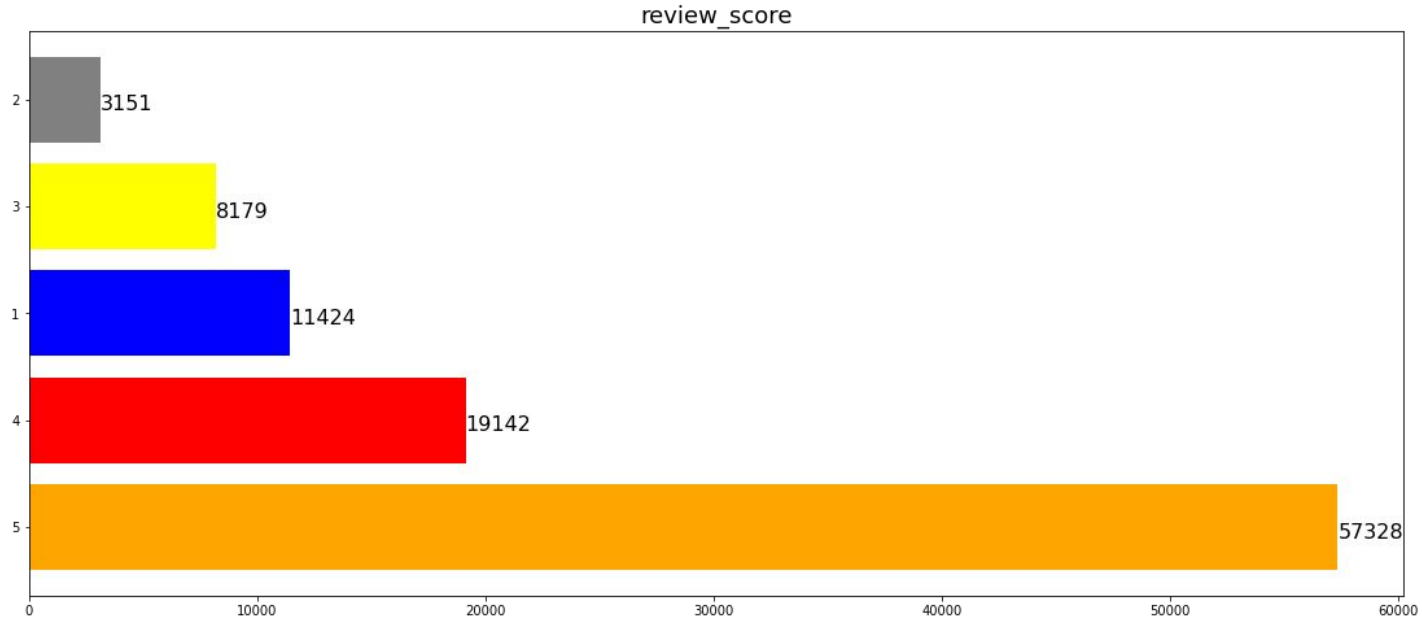
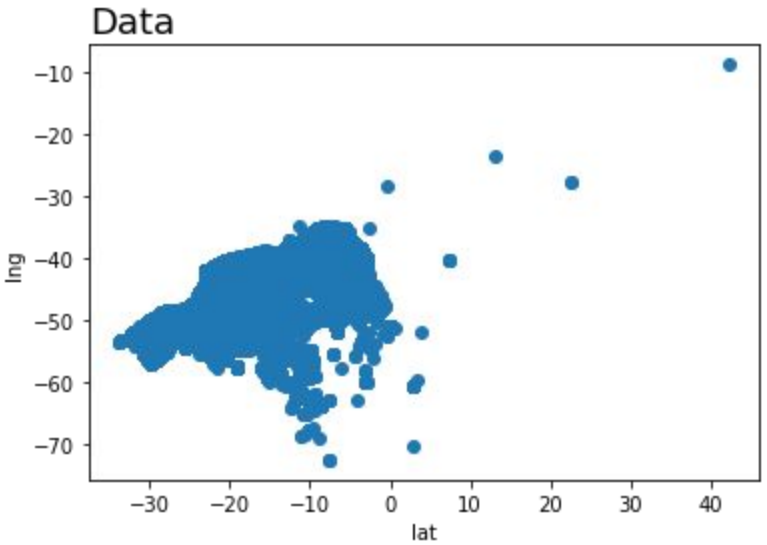
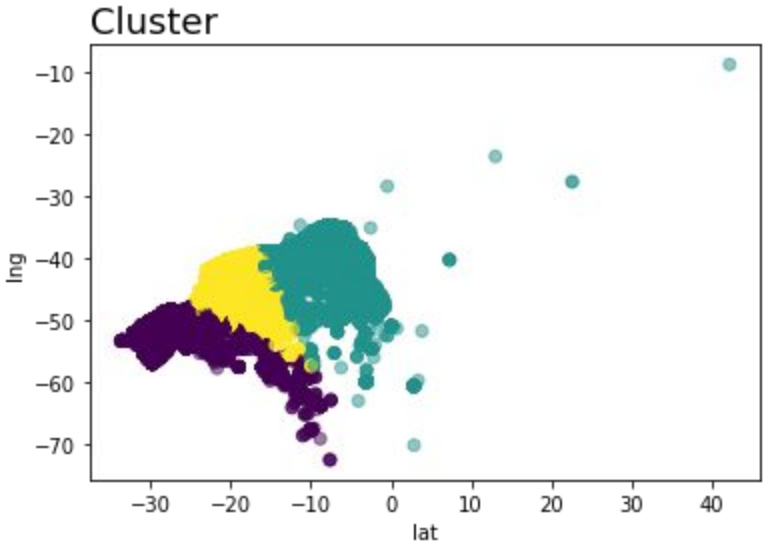


Gráfico das avaliações observadas no E-commerce. Pode-se perceber que, em uma análise geral, os produtos vendidos tiveram uma nota consideravelmente positiva.

Clientes - dados e cluster

Dispersão dos clientes pela latitude e longitude

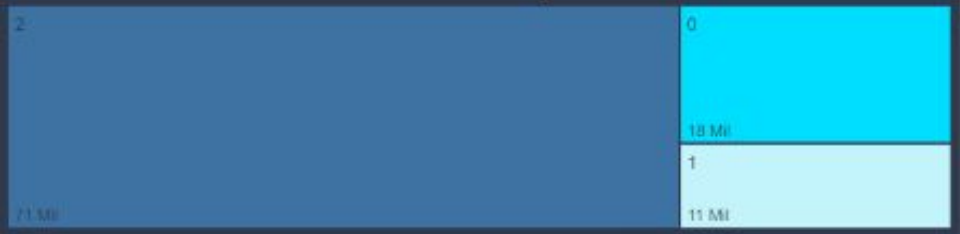




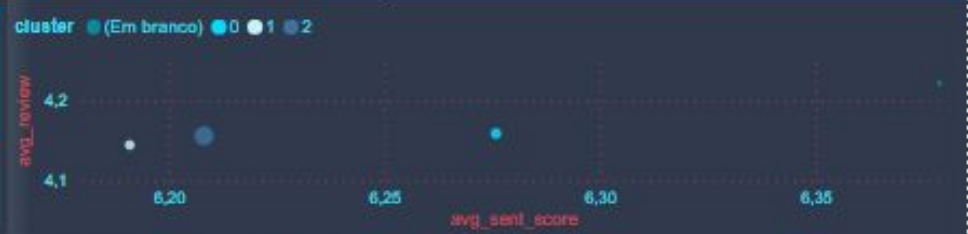
Dispersão dos clientes após a clusterização em 3 grupos distintos

Clientes Segmentados

Quantidade de clientes por clusters



Análise dos Clusters | Notas Score VS Análise de Sentimentos



Quantidade de clientes segmentados por região

Nuvem de palavras dos comentários das reviews

Segmentação dos clientes aplicado aos estados do Brasil

Linha do tempo da entrada de clientes por cluster

Dashboard e Análises

Vendedores

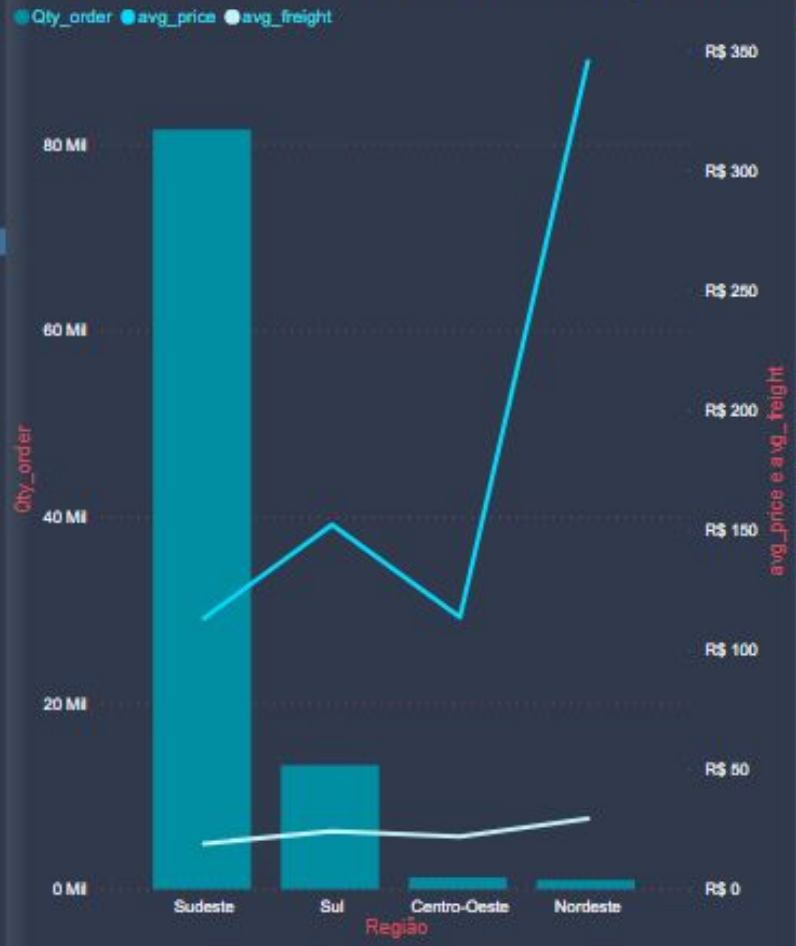
Análise dos vendores

Quantidade de vendores

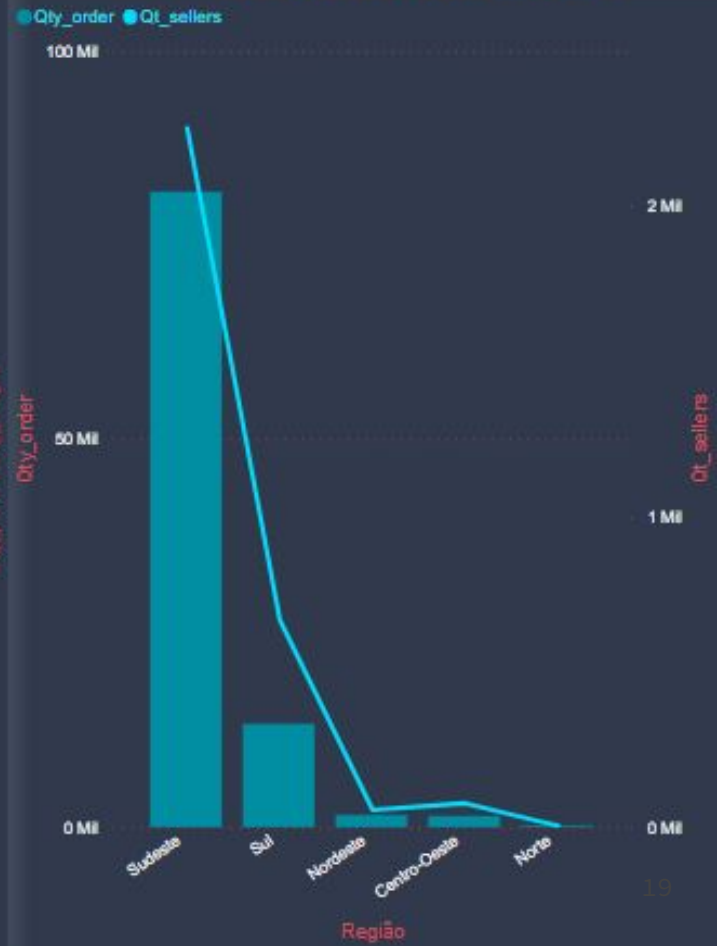
3053

Quantidade de vendedores por estado

Quantidade de vendedores por região, média de preço e frete



Quantidade pedido por regiões dos vendedores

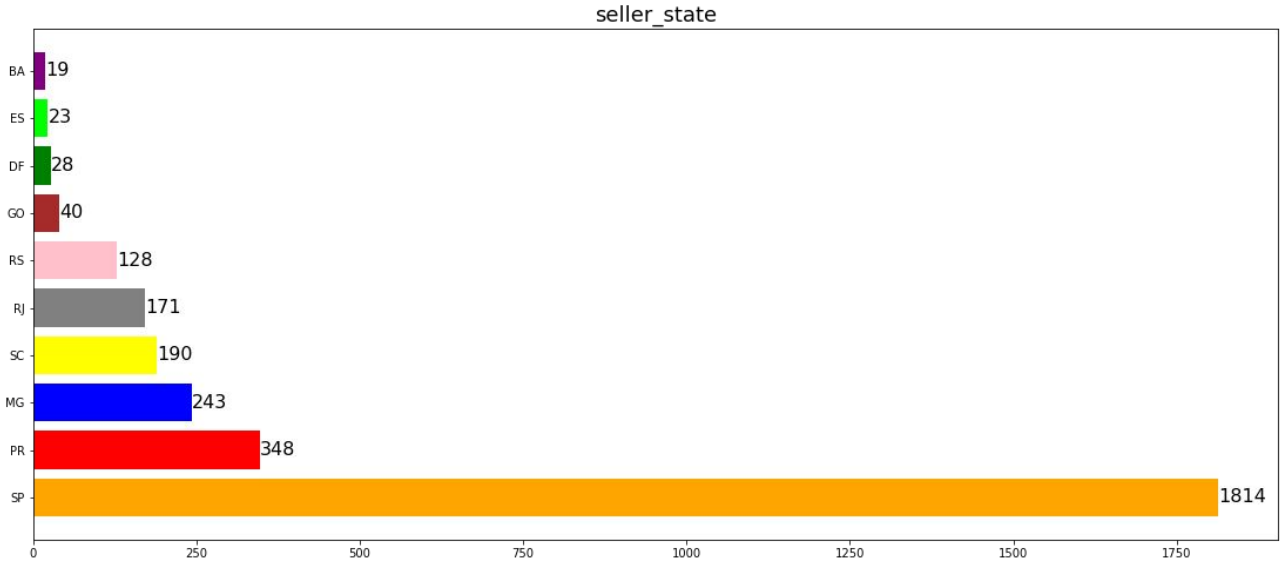


Vendedores com maior e menor valor de venda

Vendedor	Estado	Itens vendidos	Valor de venda	Frete total
4869f7a5dfa277a7dca6462dcf3b52b2	SP	1156	R$ 229.473	R$ 20.168
53243585a1d6dc2643021fd1853d8905	BA	410	R$ 222.776	R$ 13.081
4a3ca9315b744ce9f8e9374361493884	SP	1987	R$ 200.473	R$ 35.067
fa1c13f2614d7b5c4749cbc52fecda94	SP	586	R$ 194.042	R$ 10.043
7c67e1448b00f6e969d365cea6b010ab	SP	1364	R$ 187.924	R$ 51.613
7e93a43ef30c4f03f38b393420bc753a	SP	340	R$ 176.432	R$ 6.322
da8622b14eb17ae2831f4ac5b9dab84a	SP	1551	R$ 160.237	R$ 24.956
7a67c85e85bb2ce8582c35f2203ad736	SP	1171	R$ 141.746	R$ 20.903
1025f0e2d44d7041d6cf58b6550e0bfa	SP	1428	R$ 138.969	R$ 33.892
955fee9216a65b617aa5c0531780ce60	SP	1499	R$ 135.172	R$ 25.431

Vendedor	Estado	Itens vendidos	Valor de venda	Frete total
cf6f6bc4df3999b9c6440f124fb2f687	SP	1	R$4	R$ 9
77128dec4bec4878c37ab7d6169d6f26	SP	1	R$ 7	R$ 9
1fa2d3def6adfa70e58c276bb64fe5bb	SP	1	R$ 7	R$ 9
4965a7002cca77301c82d3f91b82e1a9	SP	1	R$ 8	R$ 8
ad14615bdd492b01b0d97922e87cb87f	SC	1	R$ 8	R$ 11
34aefe746cd81b7f3b23253ea28bef39	PR	1	R$ 8	R$ 15
702835e4b785b67a084280efca355756	MG	1	R$ 8	R$ 11
0f94588695d71662beec8d883ffacf09	SC	1	R$ 9	R$ 19
95cca791657aabeff15a07eb152d7841	PR	1	R$ 10	R$ 18
c18309219e789960add0b2255ca4b091	RJ	1	R$ 10	R$ 14

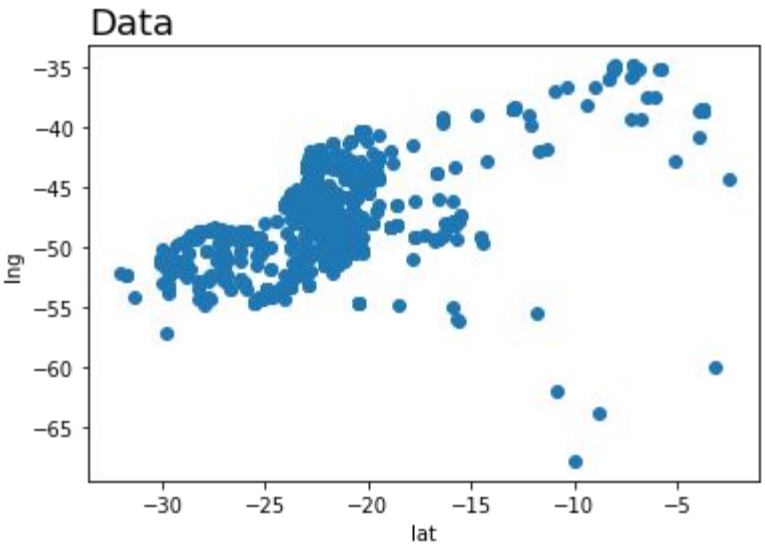
Top 10 estados com mais vendedores

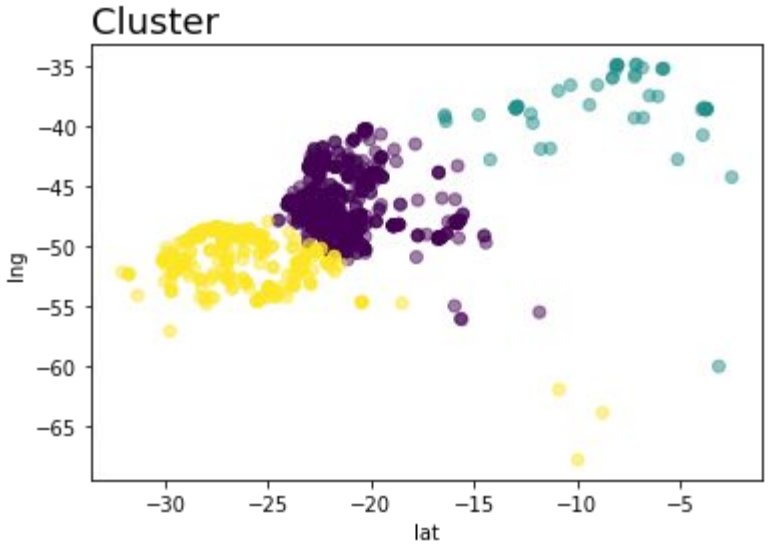


A maior concentração de vendedores se encontra no estado de São Paulo, com uma quantidade extremamente elevada de vendedores em relação aos outros estados.

Vendedores - dados e cluster

Dispersão dos vendedores pela latitude e longitude

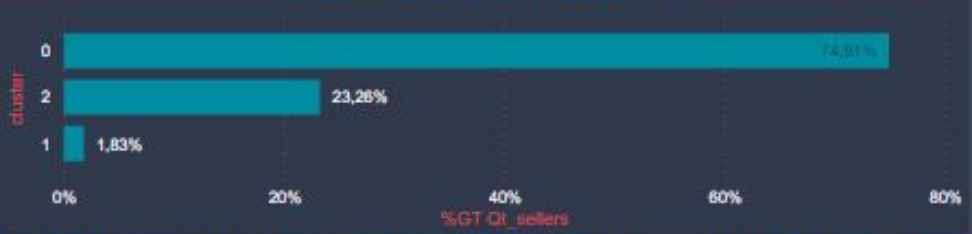




Dispersão dos vendedores após a clusterização em 3 grupos distintos

Vendedores Segmentados

Quantidade de vendedores por cluster

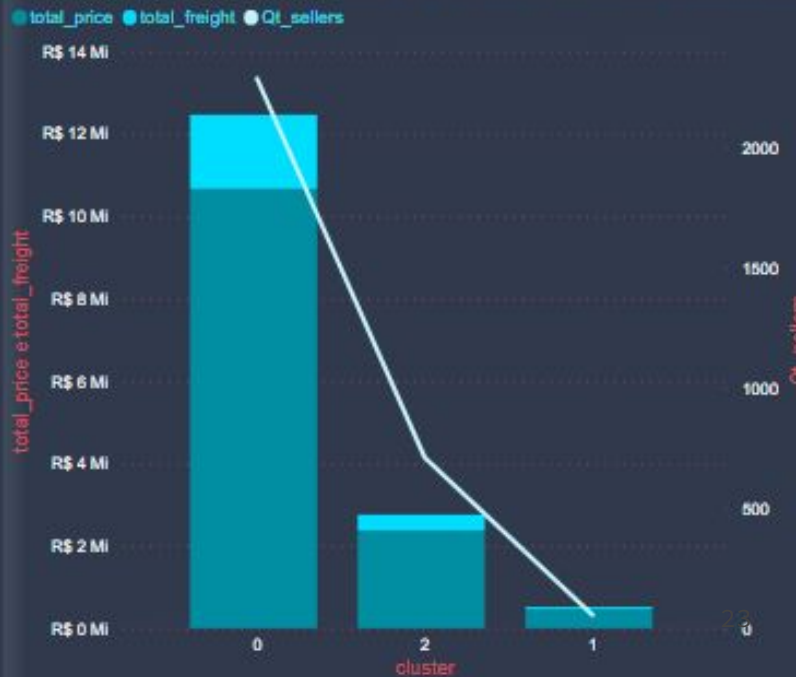


Quantidade de pedido por cluster dos vendedores

Divisão dos vendedores por estado

Segmentação dos vendedores por região, média do preço e frete

Valor total frete e preço por cluster dos clientes | Quantidade de clientes



Dashboard e Análises

Itens dos pedidos

Análise dos pedidos

99441	R$ 120,65	R$ 13.591.643,7	52546	51338	R$ 16.008.872,12
Quantidade de Pedidos	Média do preço	Preço total	Pagamentos em 1x	Pagamentos parcelados	Total pago

Quantidade pedidos por hora e média do preço por hora

Quantidade de pedidos por dia da semana, comparando a média do preço

7423	89039	R$ 2.251.909,54	R$ 19,99	11,64	22,89
Pedidos com atrasados	Entregues no prazo	Valor total do Frete	Média do valor do Frete	Média de dias para entrega	Média de dias estimados para...

Média do pagamento por tipo de pagamento

Média do preço e frete das top 5 categorias com maior média de preço

771,49	48,48	2,19	16.564,10	32949
Média de caracteres nas descrição	Média de caracteres nome	Média de fotos	Média do Volume	Quantidade de produtos

Produtos com maior e menor valor de venda

Produto	Categoria	Total	Valor de venda	Frete total
bb50f2e236e5eea0100680137654686c	beleza_saude	195	R$ 63.885	R$ 3.721
6cdd53843498f92890544667809f1595	beleza_saude	156	R$ 54.730	R$ 4.364
d6160fb7873f184099d9bc95e30376af	pcs	35	R$ 48.899	R$ 1.427
d1c427060a0f73f6b889a5c7c61f2ac4	informatica_acessorios	343	R$ 47.215	R$ 13.762
99a4788cb24856965c36a24e339b6058	cama_mesa_banho	488	R$ 43.026	R$ 8.046
3dd2a17168ec895c781a9191c1e95ad7	informatica_acessorios	274	R$ 41.083	R$ 7.130
25c38557cf793876c5abdd5931f922db	bebes	38	R$ 38.907	R$ 1.405
5f504b3a1c75b73d6151be81eb05bdc9	cool_stuff	63	R$ 37.734	R$ 3.992
53b36df67ebb7c41585e8d54d6772e08	relogios_presentes	323	R$ 37.683	R$ 2.275
aca2eb7d00ea1a7b8ebd4e68314663af	moveis_decoracao	527	R$ 37.609	R$ 7.212

Produto	Categoria	Total	Valor de venda	Frete total
310dc32058903b6416c71faff132df9e	papelaria	1	R$ 2	R$ 8
46fce52cef5caa7cc225a5531c946c8b	beleza_saude	1	R$ 2	R$ 7
2e8316b31db34314f393806fd7b6e185	papelaria	1	R$ 3	R$ 12
8a3254bee785a526d548a81a9bc3c9be	construcao_ferramentas_construcao	3	R$ 3	R$ 59
680cc8535be7cc69544238c1d6a83fe8	pet_shop	1	R$ 3	R$ 9
836c4b48c2b383bb38bb5788f828c596	fashion_underwear_e_moda_praia	1	R$ 4	R$ 15
c2fb26742f8484dbfe9a8d70bdc54025	informatica_acessorios	1	R$ 4	R$ 15
66389c9df136a25c8f131757ce3a6967	construcao_ferramentas_construcao	1	R$ 4	R$ 13
ba82e510acd9a0fe69a44cafea53f9aa	papelaria	1	R$ 4	R$ 12
eee2fb3dceb9ffd8a99dd4bc4b7e860a	informatica_acessorios	1	R$ 4	R$ 12

Apriori das compras

Utilizando supp min de 0.01 e conf min de 0.01

Apriori por Produto:

```
    lhs                                      rhs                                      support   confidence coverage  lift     count
{e53e557d5a159f5aa2c5e995dfdf244b} => {36f60d45225e60c7da4558b070ce4b60} 0.0105068 0.8947368  0.01174289 60.32018 34
{36f60d45225e60c7da4558b070ce4b60} => {e53e557d5a159f5aa2c5e995dfdf244b} 0.0105068 0.7083333  0.01483313 60.32018 34
```

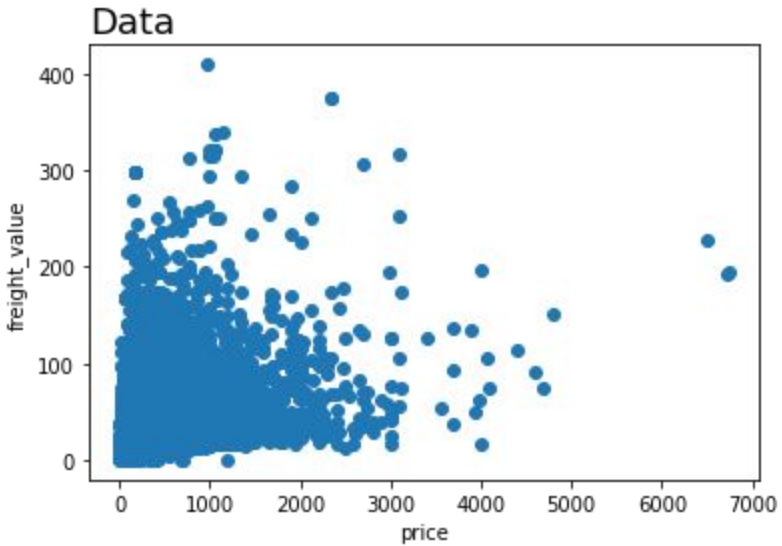
Apriori por categoria de produto:

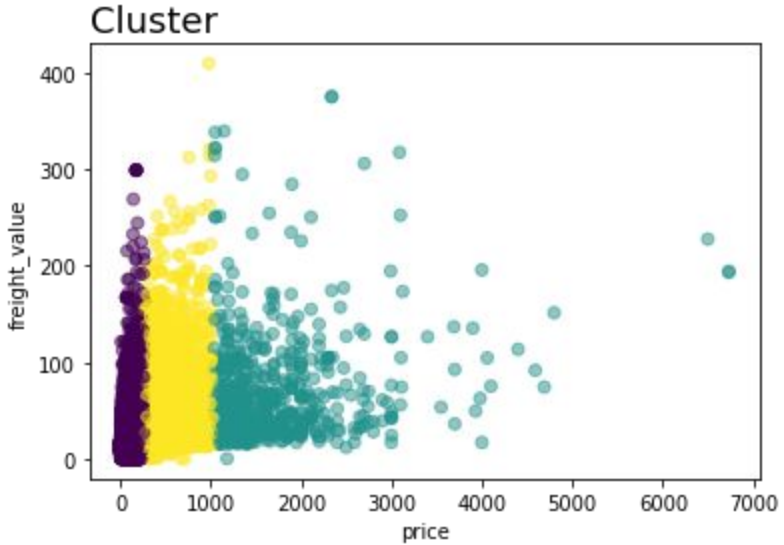
```
    lhs                  rhs                support    confidence coverage   lift  count
{casa_conforto}    => {cama_mesa_banho}  0.01328801 0.7413793  0.01792336 2.9988793 43
{cama_mesa_banho}  => {casa_conforto}    0.01328801 0.0537500  0.24721879 2.9988793 43
{moveis_decoracao} => {cama_mesa_banho}  0.02163164 0.1580135  0.13689740 0.6391648 70
{cama_mesa_banho}  => {moveis_decoracao} 0.02163164 0.0875000  0.24721879 0.6391648 70
```

O suporte das regras obtidas são muito baixos, variando apenas entre 1% e 2.2%.
A confiança tem uma variação maior, transitando de um mínimo de 5.3% e chegando até um máximo de 89.4%

Itens do pedido - dados e cluster

Dispersão dos itens dos pedidos pelo preço e frete

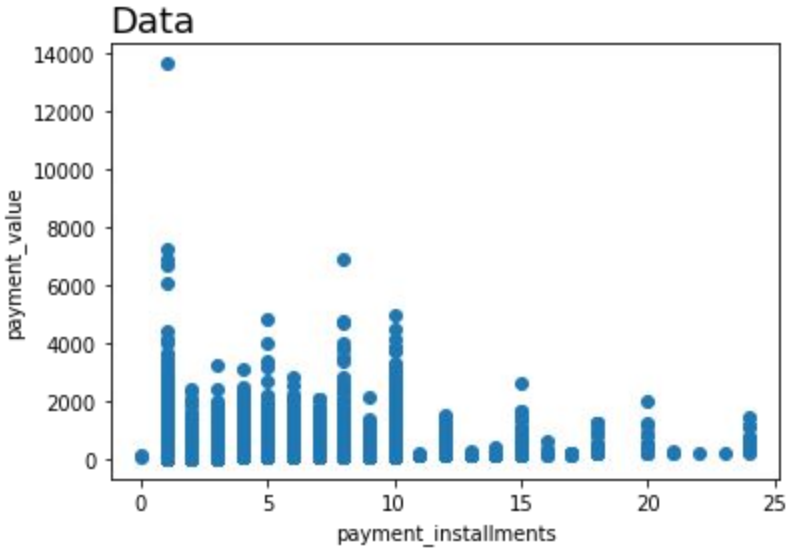


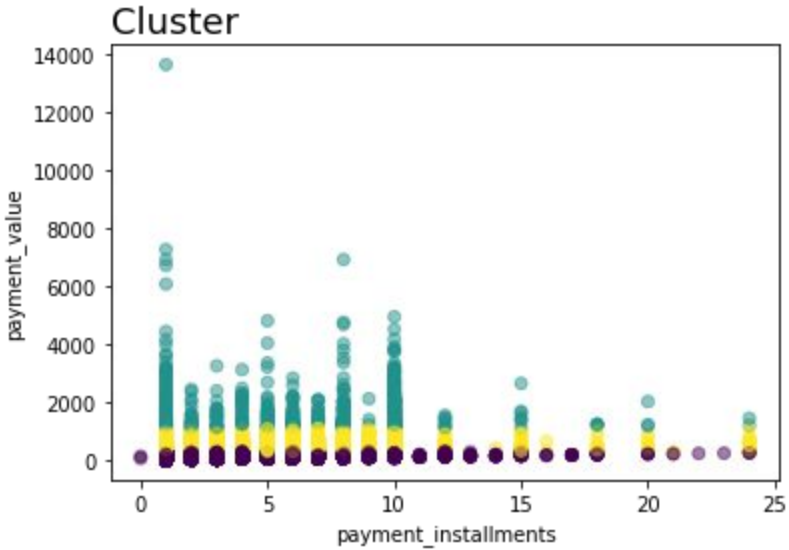


Dispersão dos itens dos pedidos após a clusterização em 3 grupos distintos

Pagamentos de pedidos - dados e cluster

Dispersão dos pagamentos pelo valor e parcelas

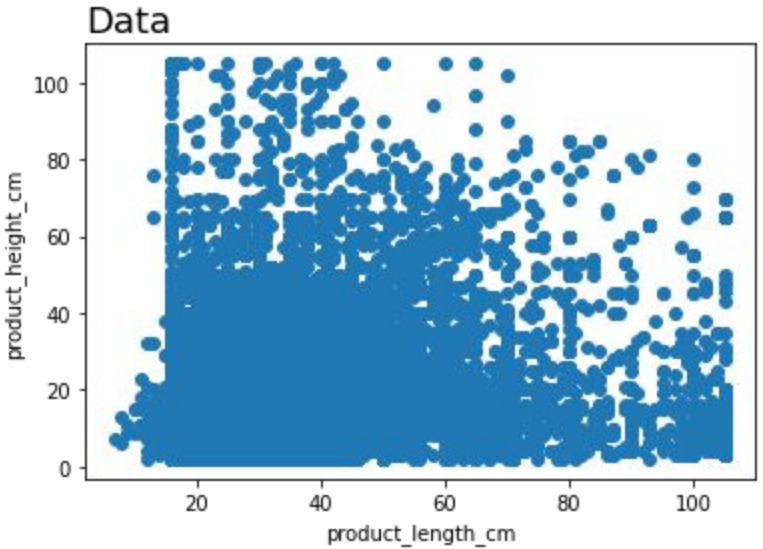


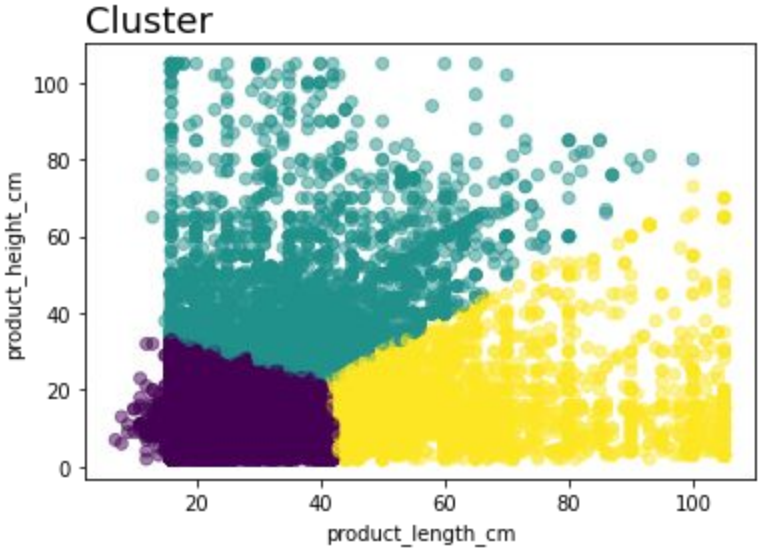


Dispersão dos pagamentos após a clusterização em 3 grupos distintos

Produtos - dados e cluster

Dispersão dos produtos pelo comprimento e altura





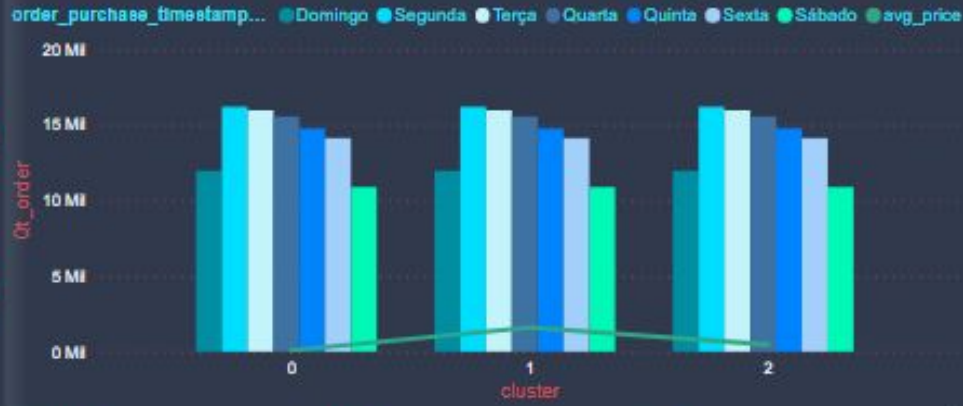
Dispersão dos produtos após a clusterização em 3 grupos distintos

Pedidos segmentados

99441	R$ 120,65	R$ 13.591.643,7	52546	51338	R$ 16.008.872,12
Quantidade de Pedidos	Média do preço	Preço total	Pagamentos em 1x	Pagamentos parcelados	Total pago

Quantidade de pedidos e média preço por hora divido por cluster

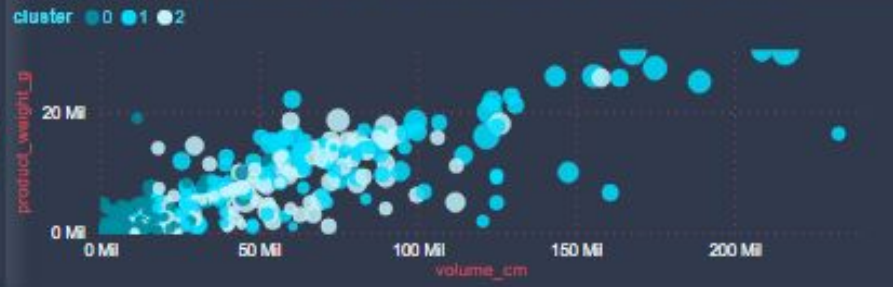
Quantidade de pedidos por dia da semana | média do preço | Clusterizado



7423	89039	R$ 2.251.909,54	R$ 19,99	11,64	22,89
Pedidos com atrasados	Entregues no prazo	Valor total do Frete	Média do valor do Frete	Média de dias para entrega	Média de dias estimados para...

Média do pagamento por cluster comparando o total do pagamento

Peso (g) por volume (cm) do produtos, comparando por cluster



771,49	48,48	2,19	16.564,10	32949
Média de caracteres nas descrição	Média de caracteres nome	Média de fotos	Média do Volume	Quantidade de produtos

“Sem dados você é apenas mais uma pessoa com uma opinião.”

W. Edwards Deming

“Os erros causados por dados inadequados são muito menores do que aqueles devido à falta total de dados.”

Charles Babbage

Obrigado!