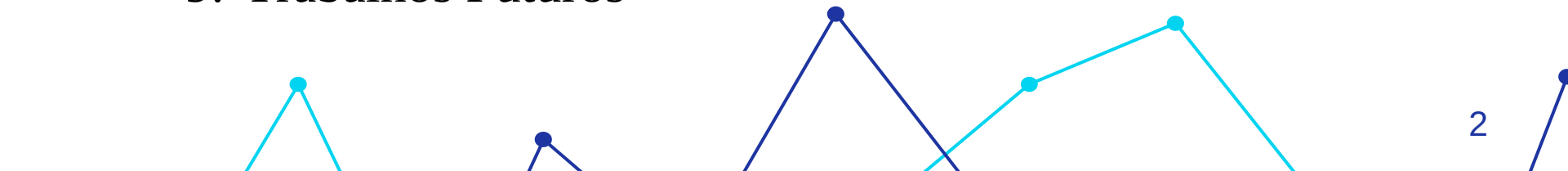
APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA IDENTIFICAR ASPECTOS QUE INFLUENCIAM NO PROCESSO DE VENDA E ENTREGA DOS PRODUTOS DE UM E-COMMERCE

Acadêmico: Herikc Brecher
Orientador: Rafael Ballottin Martins

Sumário

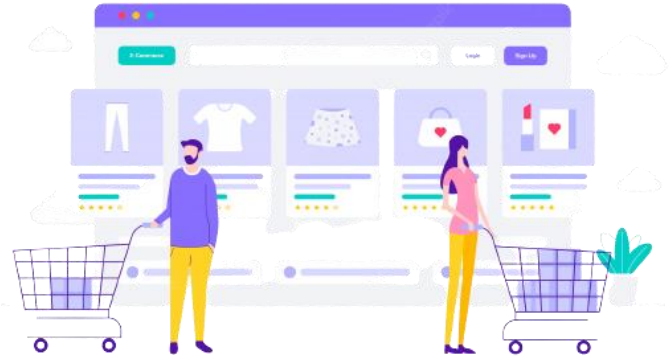


1. Contextualização

- Processo de KDD:
  - Analise Exploratória;
  - Pré-Processamento;
  - Mineração de Dados;
  - Avaliação dos resultados.

1.1 Problema de Pesquisa

- ⬆ Geração de dados;
- ⬆ Complexidade dos dados;
- Dados analisados superficialmente;
- Necessidade de analisar os dados.



1.2 Solução

- Base de dados publica;
- Processo de KDD:
  - Analise de dados;
  - Pré-Processamento;
  - Mineração de dados;
  - Interpretação e avaliação dos resultados.

2. Objetivo Geral

Identificar aspectos que influenciam no processo de avaliação dos pedidos, desempenho das vendas e entrega.

2.1 Objetivos Específicos

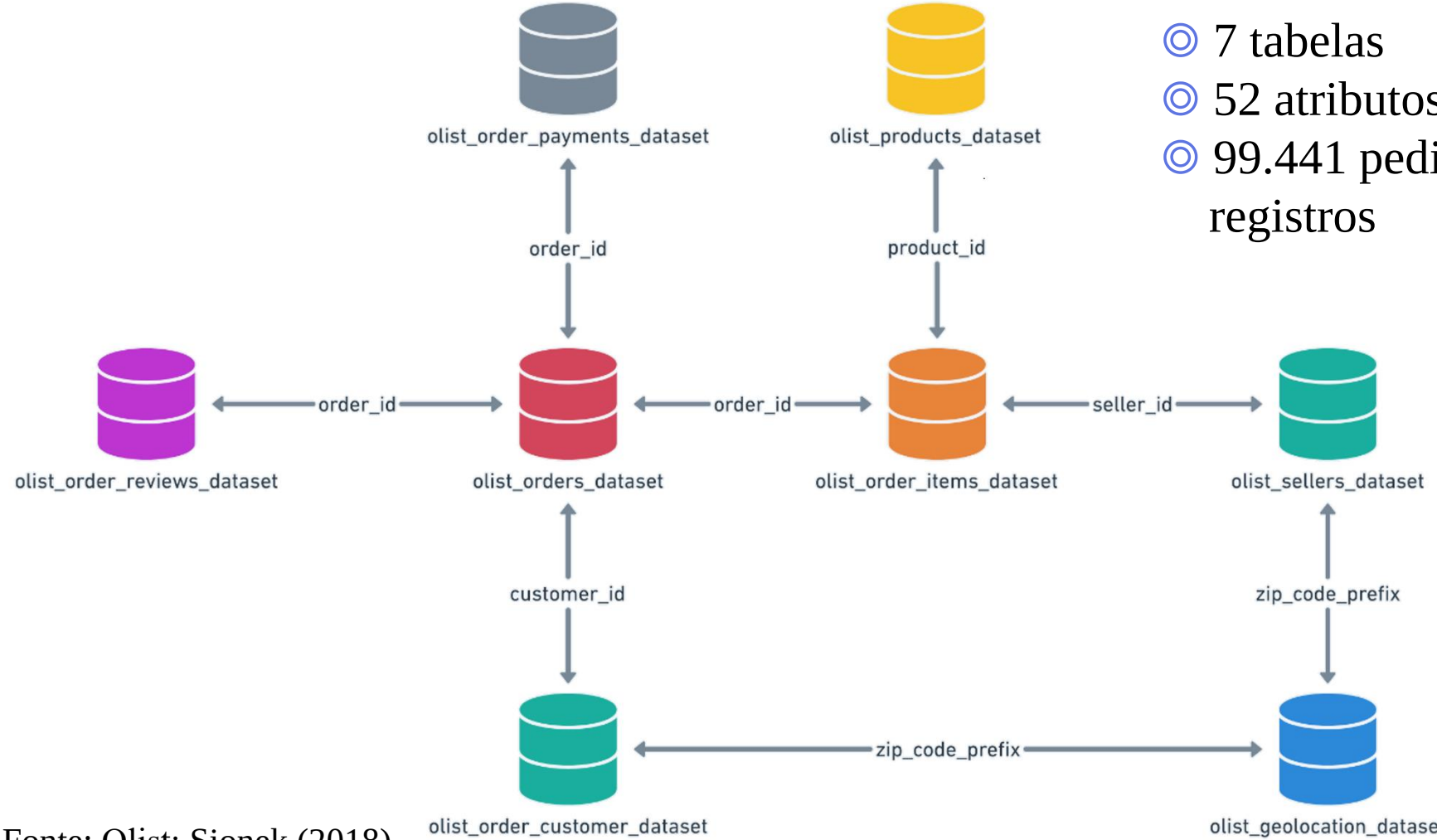
◎ Definir os atributos mais relevantes através de uma análise exploratória dos dados;

◎ Identificar os principais aspectos que influenciam a avaliação de pedidos, desempenho das vendas e entrega da empresa;

◎ Identificar e definir algoritmos para regressão e clusterização;

◎ Avaliar os resultados a fim de descobrir o nível de confiança das informações obtidas.

PROJETO



- 7 tabelas
- 52 atributos
- 99.441 pedidos / registros

Fonte: Olist; Sionek (2018)

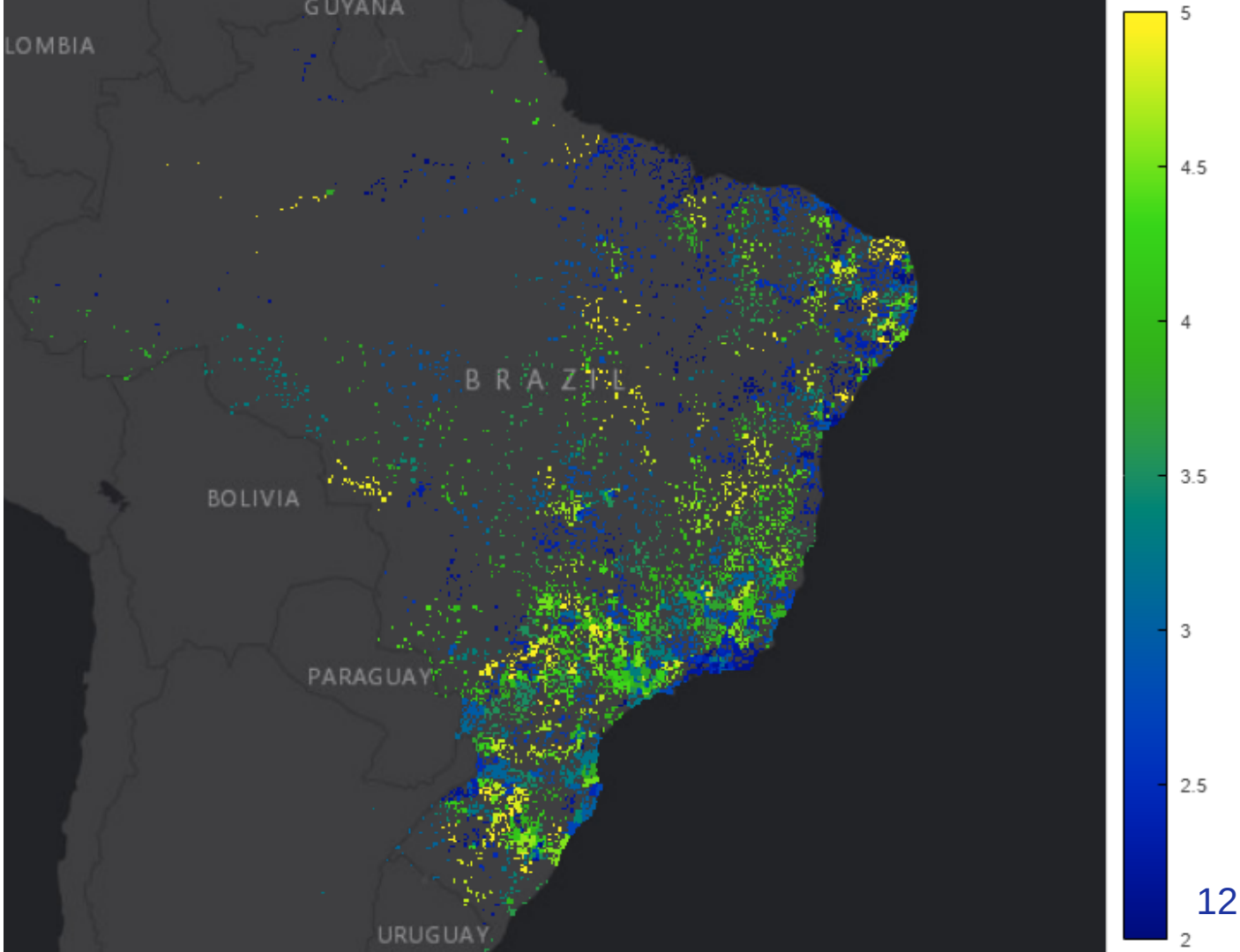
3. Projeto

Atributo	Valores Únicos
Clientes	96.096
Pedidos	99.441
Avaliações	98.410
Lojas	3.095
Produtos	71

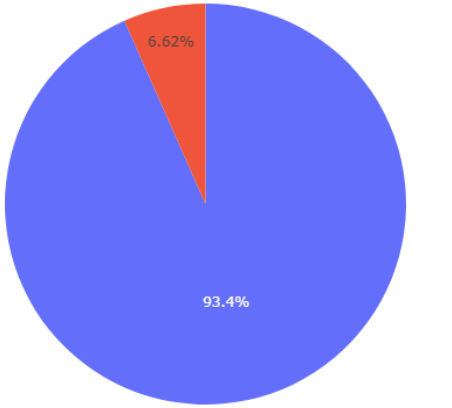
Fonte: Olist; Sionek (2018) – Adaptado pelo autor.

ANALISE EXPLORATÓRIA

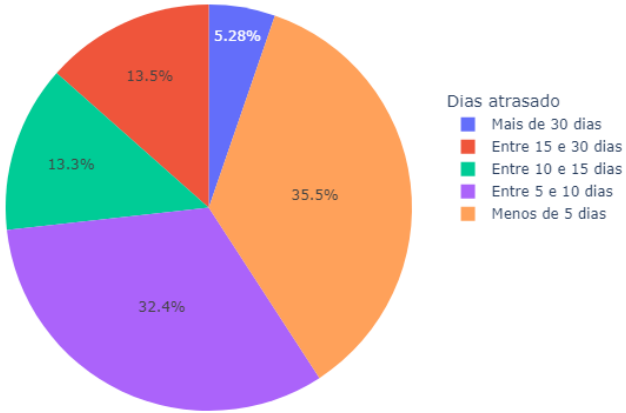
Review médio por região no Brasil



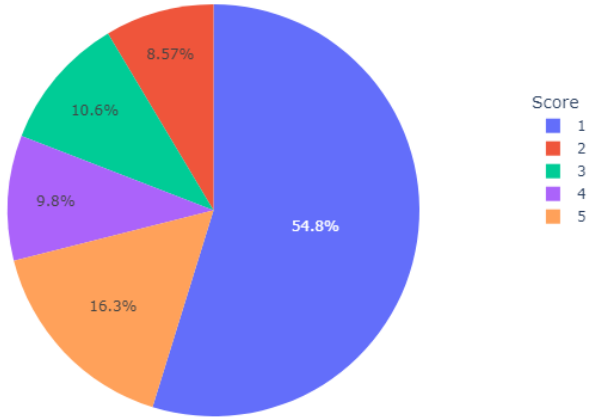
% de Pedidos Entregues no Período



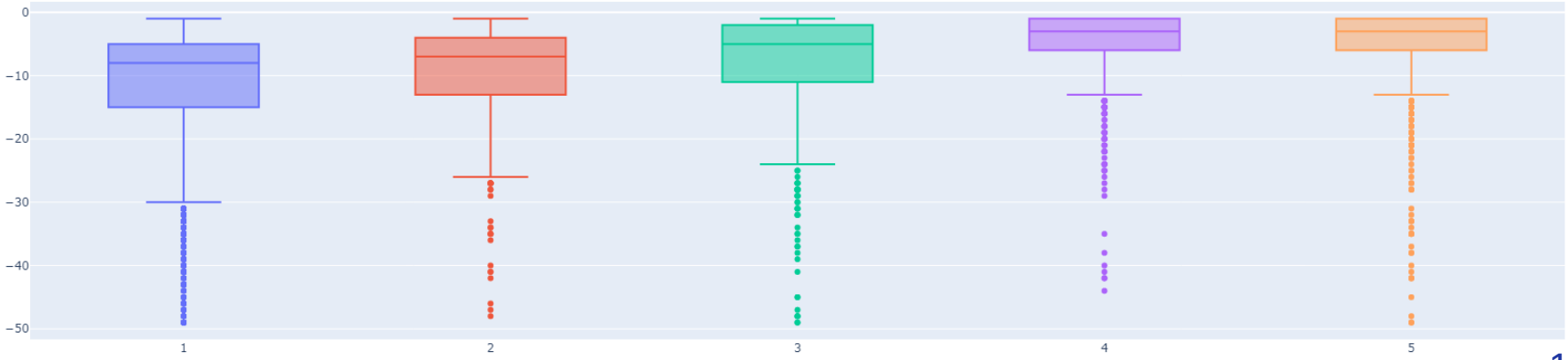
% Por Categoria de Dias de Atraso

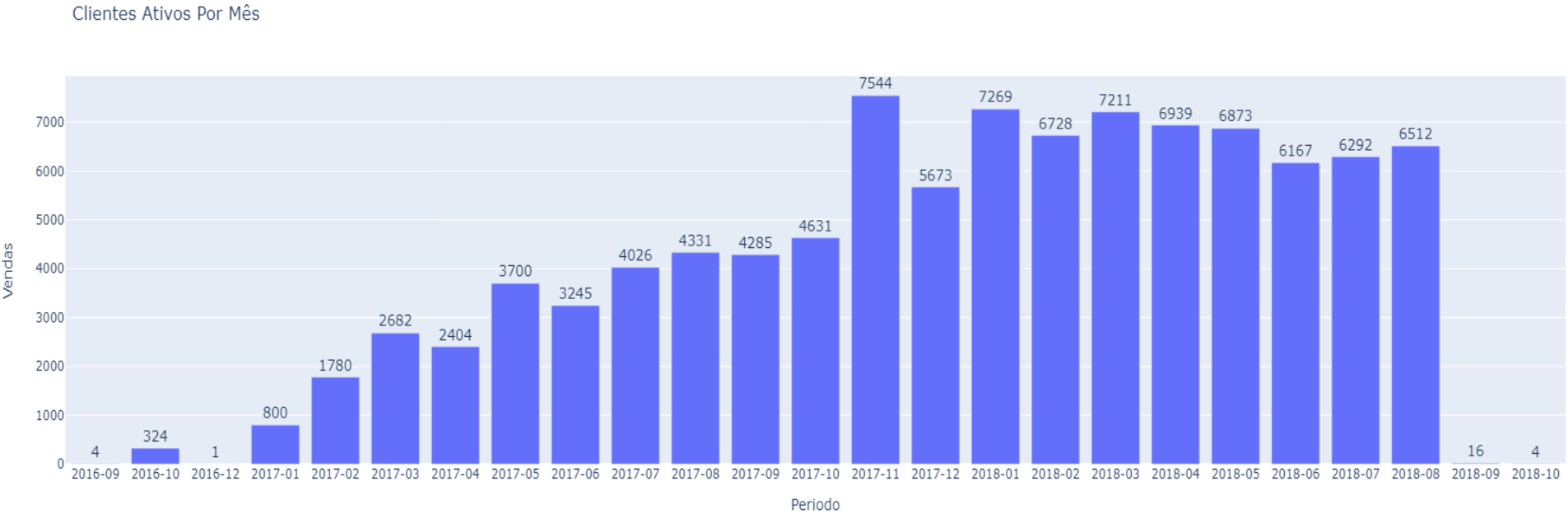


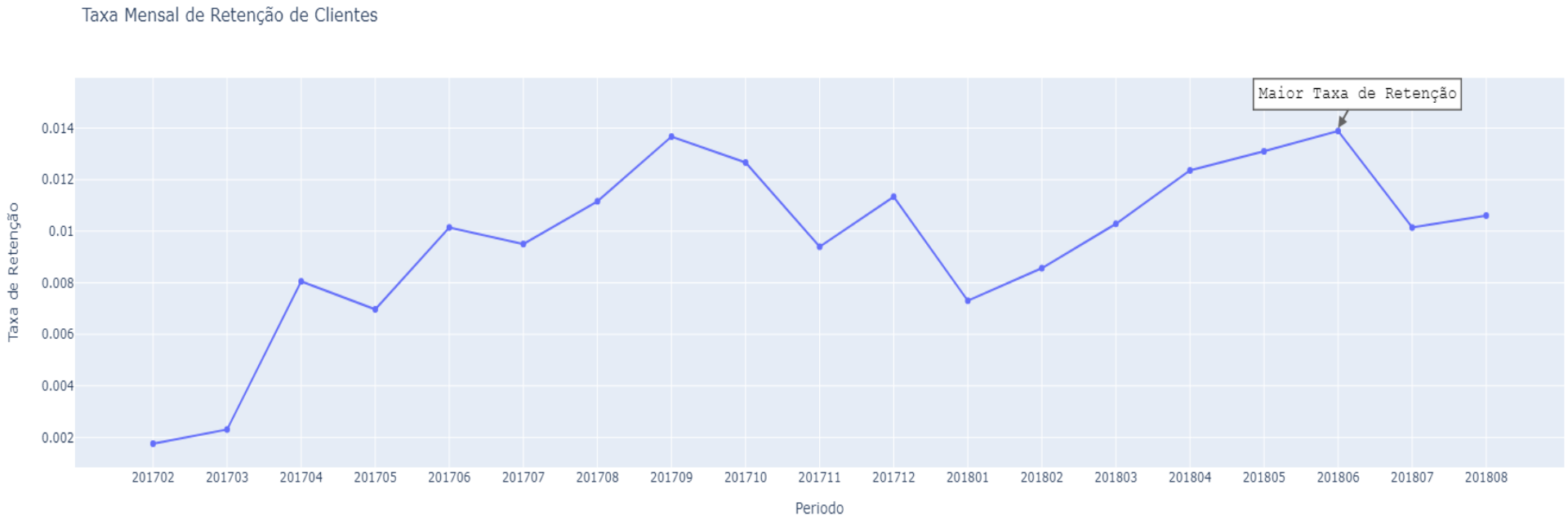
% de Pedidos Atrasados por Score

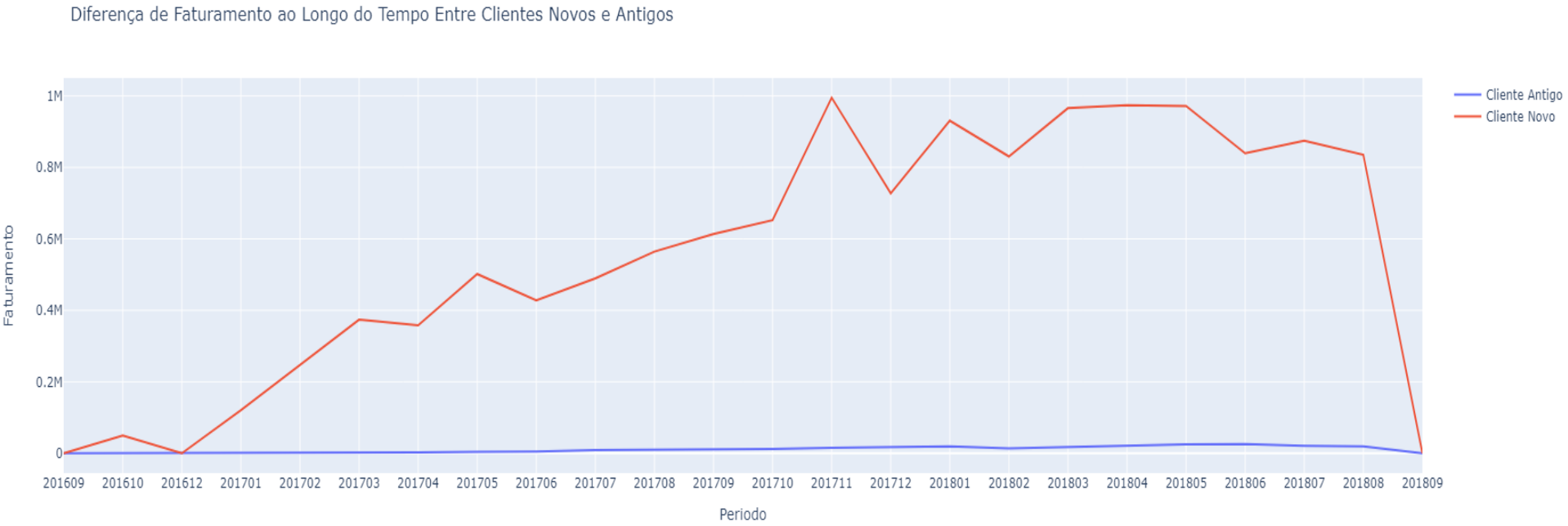


Dias de Atraso por Score




Clientes Ativos Por Mês
Vendas
Periodo
0
1000
2000
3000
4000
5000
6000
7000
4
324
1
800
1780
2682
2404
3700
3245
4026
4331
4285
4631
7544
5673
7269
6728
7211
6939
6873
6167
6292
6512
16
4
2016-09
2016-10
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09
2018-10


Taxa Mensal de Retenção de Clientes
Maior Taxa de Retenção
Taxa de Retenção
0.014
0.012
0.01
0.008
0.006
0.004
0.002
201702
201703
201704
201705
201706
201707
201708
201709
201710
201711
201712
201801
201802
201803
201804
201805
201806
201807
201808
Periodo


Diferença de Faturamento ao Longo do Tempo Entre Clientes Novos e Antigos
Cliente Antigo
Cliente Novo
Faturamento
1M
0.8M
0.6M
0.4M
0.2M
0
201609
201610
201612
201701
201702
201703
201704
201705
201706
201707
201708
201709
201710
201711
201712
201801
201802
201803
201804
201805
201806
201807
201808
201809
Periodo

PROCESSAMENTO DOS DADOS

Teste de hipótese por meio de Grubs para detecção de outliers.

Atributo	Possui Outliers
freight_value	Sim
payment_value	Sim
quantity	Sim
diff_delivery_and_estimate	Sim
diff_delivery_and_purchase	Sim

Aplicação de *Label Encoder*.

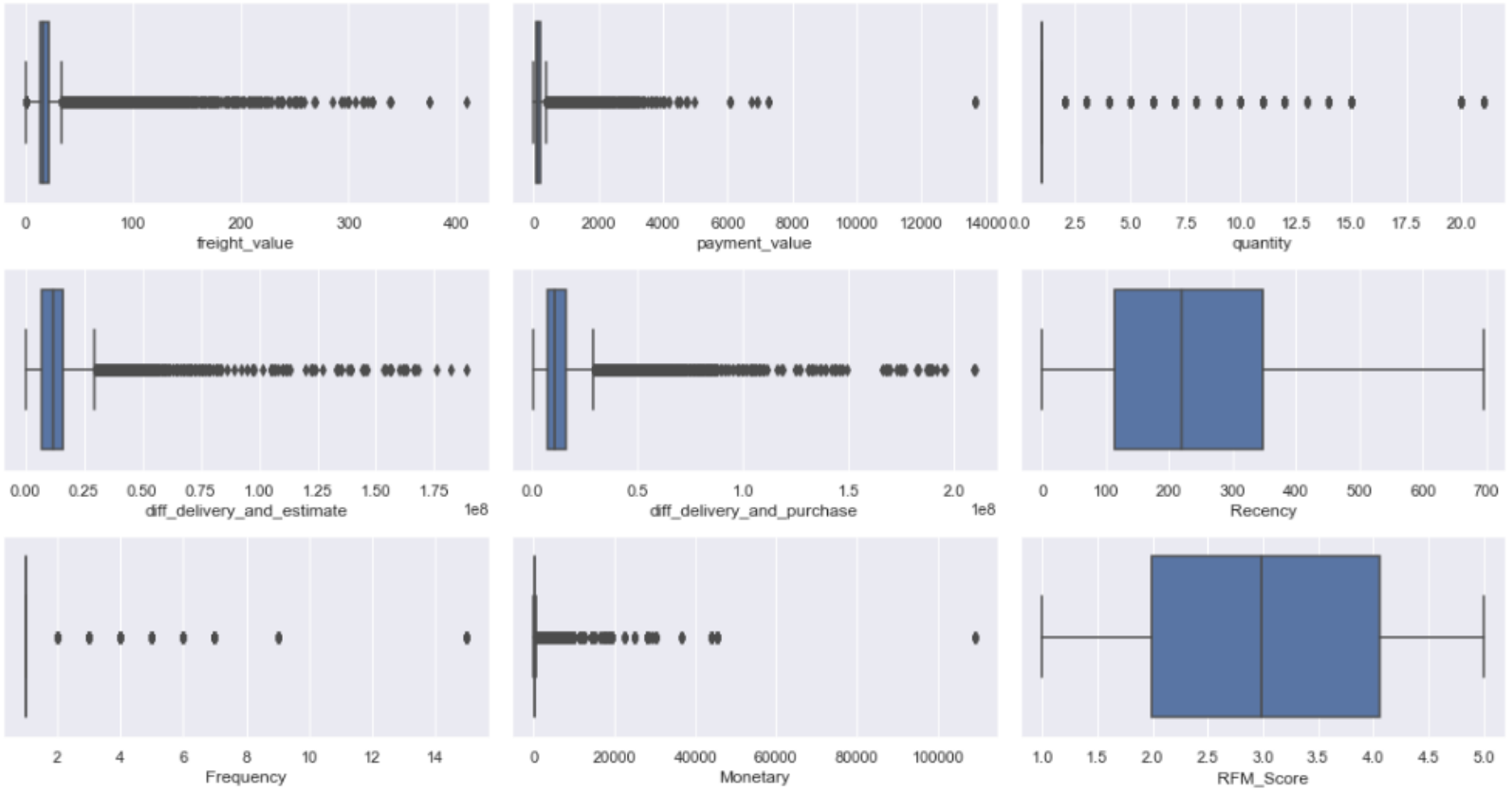
Antes da aplicação:

freight_value	payment_type	payment_installments	payment_value	review_score	quantity	diff_delivery_and_estimate	diff_delivery_and_purchase	customer_state	customer_region
13.29	credit_card	2	72.19	5	1	9.0	7.0	RJ	Sudeste
19.93	credit_card	3	259.83	4	1	3.0	16.0	SP	Sudeste
17.87	credit_card	5	216.87	5	1	14.0	8.0	MG	Sudeste
12.79	credit_card	2	25.78	4	1	6.0	6.0	SP	Sudeste
18.14	credit_card	3	218.04	5	1	16.0	25.0	SP	Sudeste

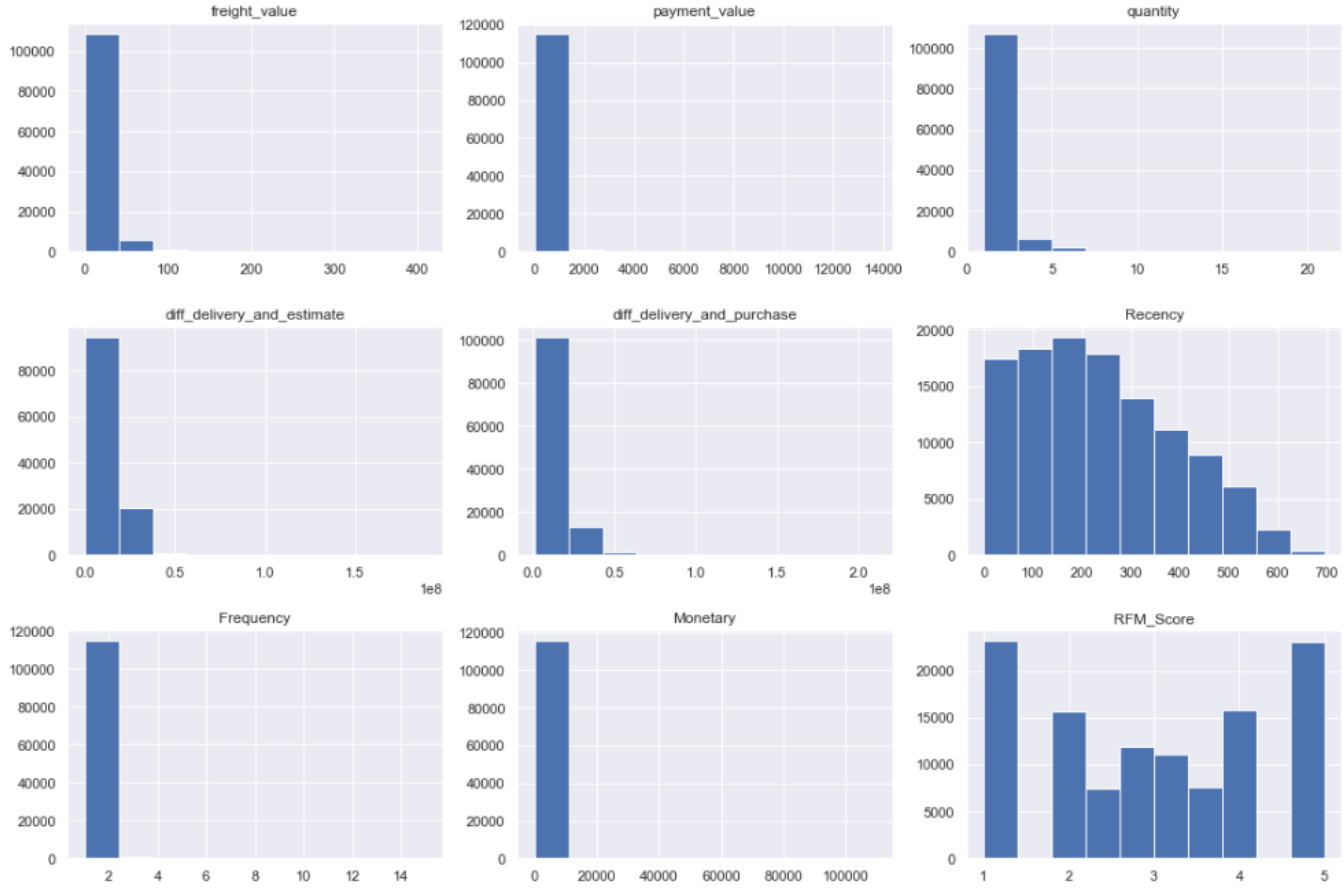
Depois da aplicação:

freight_value	payment_type	payment_installments	payment_value	review_score	quantity	diff_delivery_and_estimate	diff_delivery_and_purchase	customer_state	customer_region
13.29	1	2	72.19	5	1	9.0	7.0	18	3
19.93	1	3	259.83	4	1	3.0	16.0	25	3
17.87	1	5	216.87	5	1	14.0	8.0	10	3
12.79	1	2	25.78	4	1	6.0	6.0	25	3
18.14	1	3	218.04	5	1	16.0	25.0	25	3

Boxplot dos atributos quantitativos:



Histograma dos atributos quantitativos:



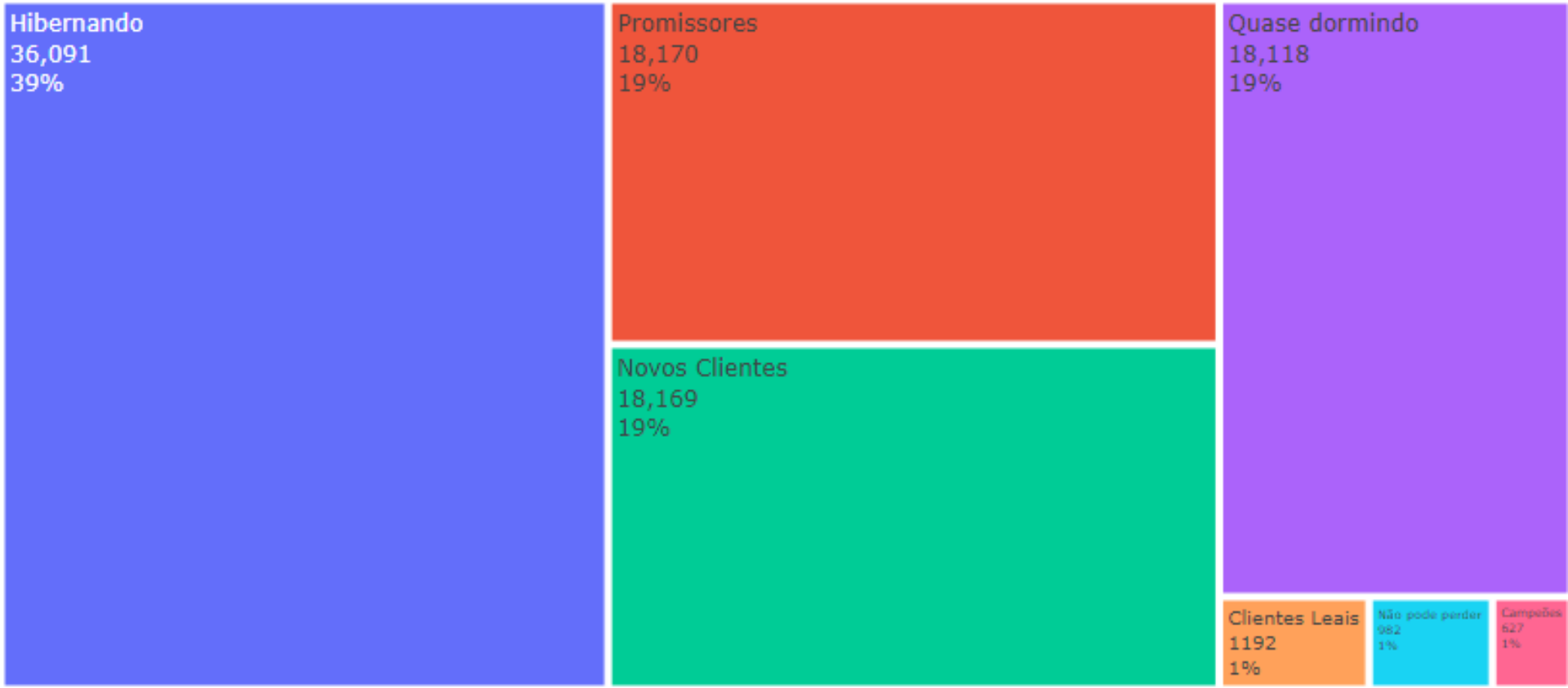
Simetria dos atributos quantitativos:

Variável	Skewness	Kurtosis
freight_value	5.558939	58.300861
payment_value	14.448540	530.034479
quantity	6.331864	68.366469
diff_delivery_and_estimate	2.846043	32.528515
diff_delivery_and_purchase	3.912217	40.794495
Recency	0.447031	-0.656545
Frequency	11.008832	258.347088
Monetary	29.531411	1432.134347
RFM_Score	-0.003545	-1.271986

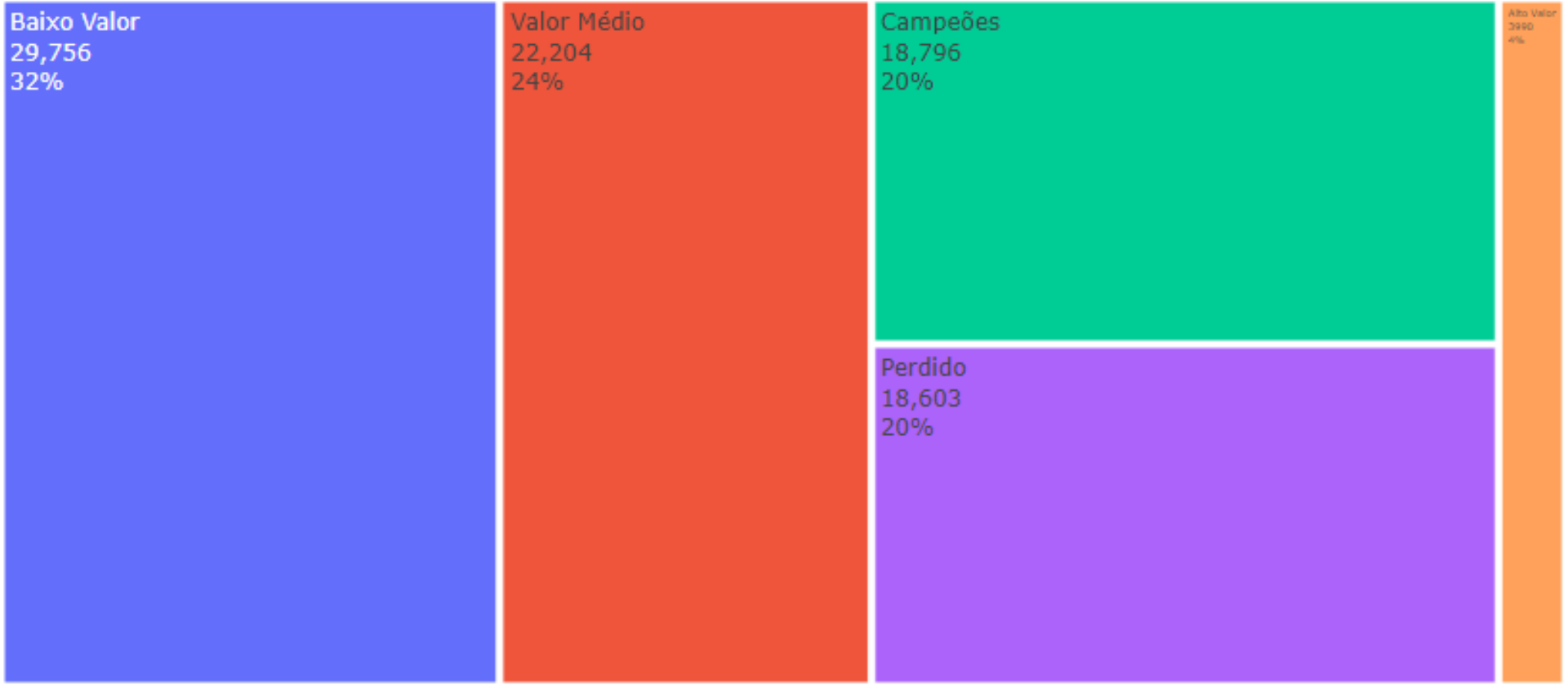
Clusterização

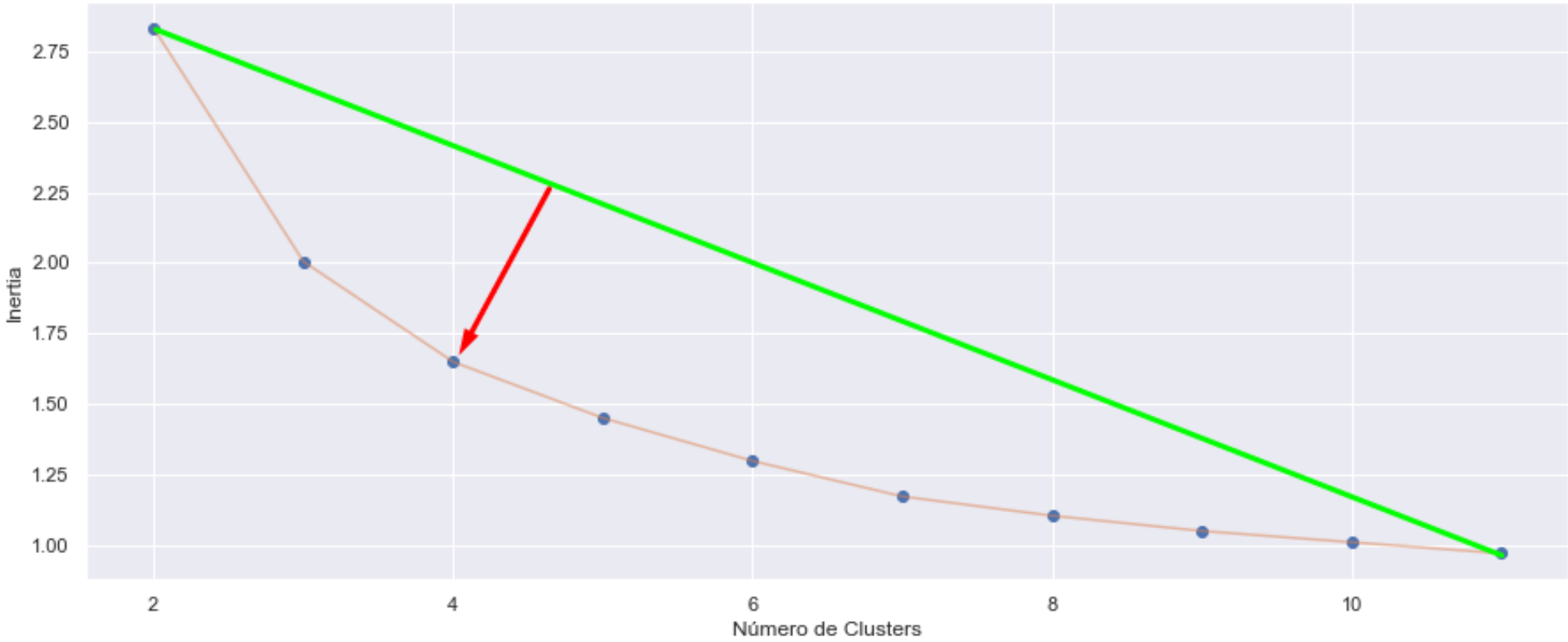
Categoria	Descrição
Recência	Quão recentemente o cliente fez uma transação. Por exemplo, o cliente realizou a sua última compra a 30 dias.
Frequência	Quão frequente o cliente realizar um pedido. Por exemplo, nos últimos 12 meses foram realizadas 4 compras.
Monetário	Quanto o cliente gastou ao total nos pedidos.

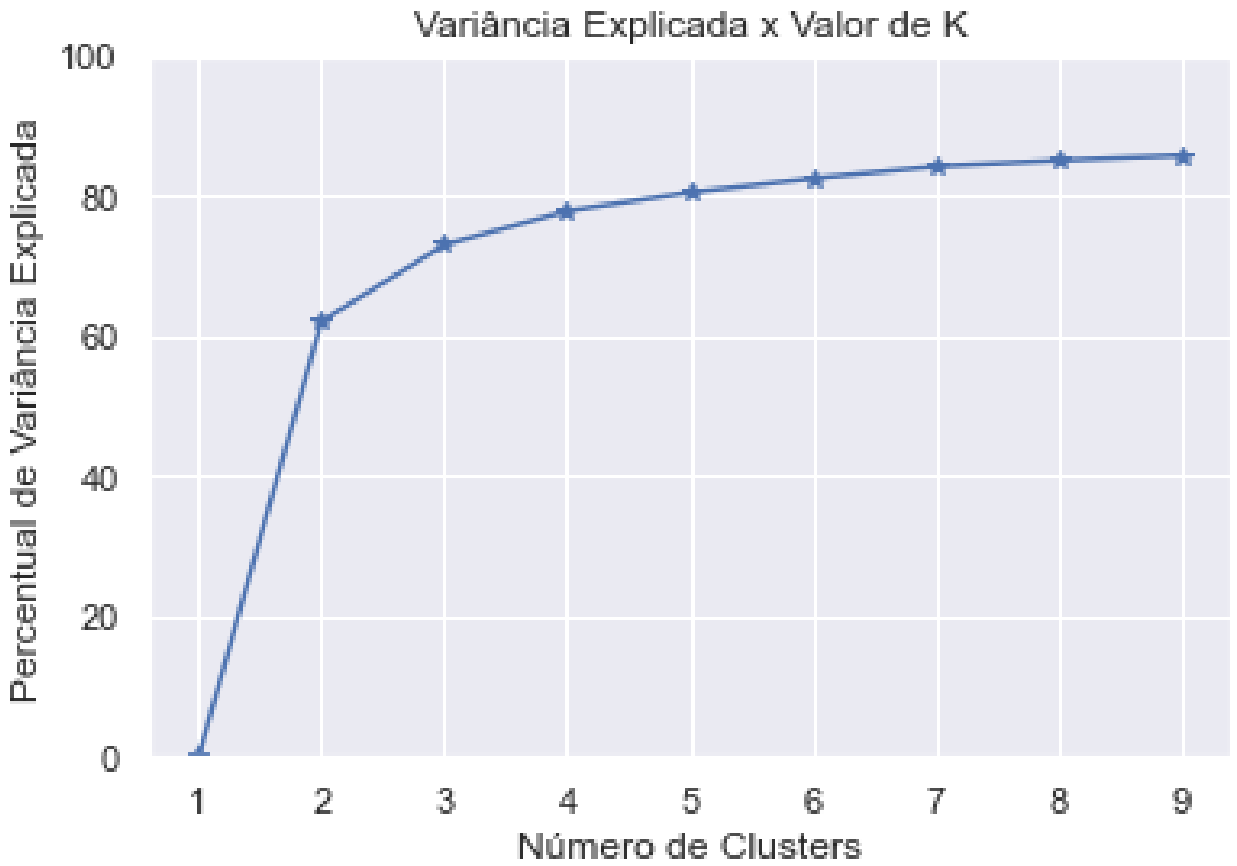
Analise de RFM por Recência e Frequência



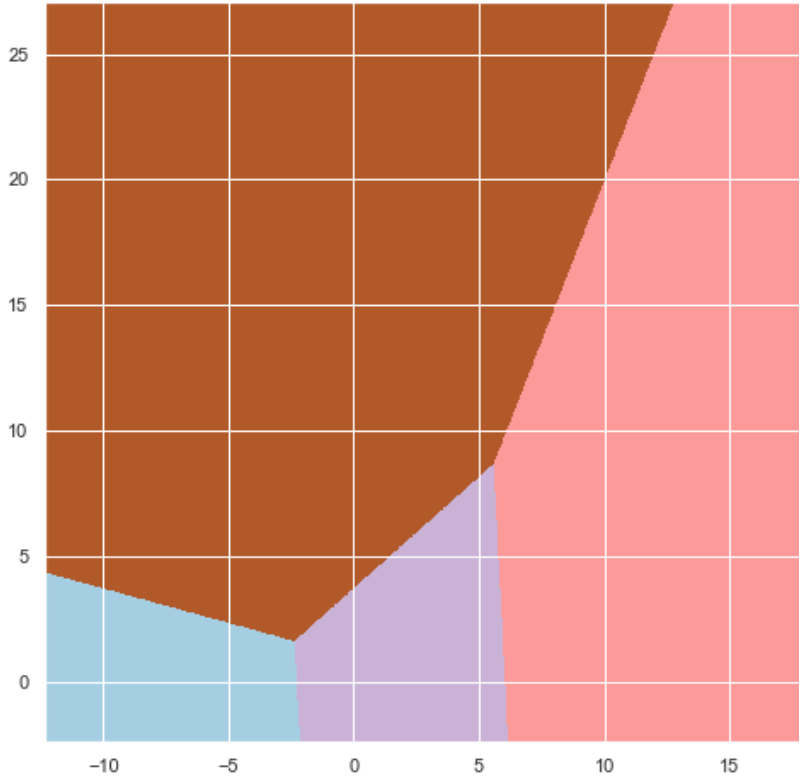
Analise de RFM por Recência, Frequência e Valor Monetário



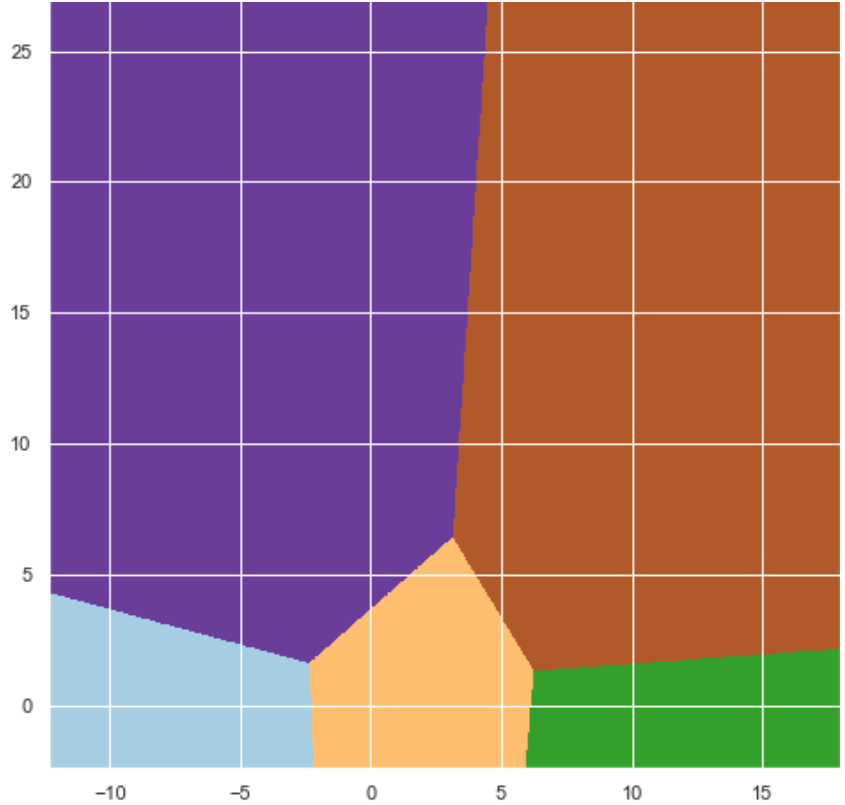

Inertia
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
2
4
6
8
10
Número de Clusters


Variância Explicada x Valor de K
Percentual de Variância Explicada
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
Número de Clusters

Meshgrid de 4 clusters com PCA

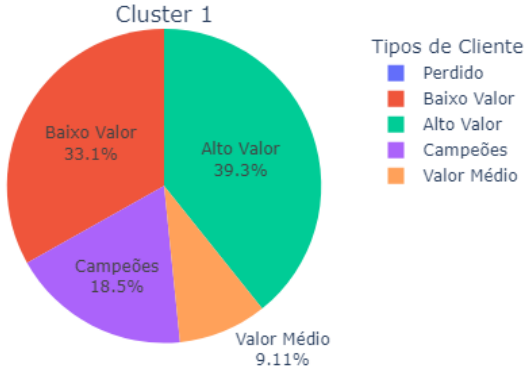


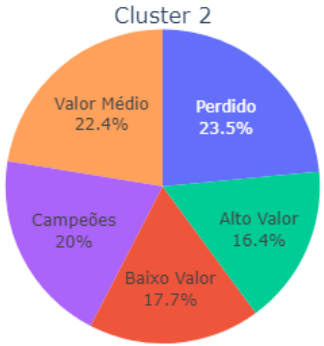
Meshgrid de 5 clusters com PCA

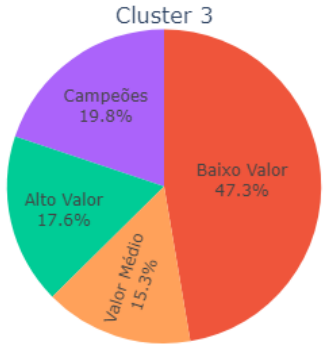


K-Clusters	Silhouette Score com PCA	Silhouette Score sem PCA
4	0.6094	0.4456
5	0.6025	0.4308
6	0.6114	0.4057
8	0.5799	0.3219

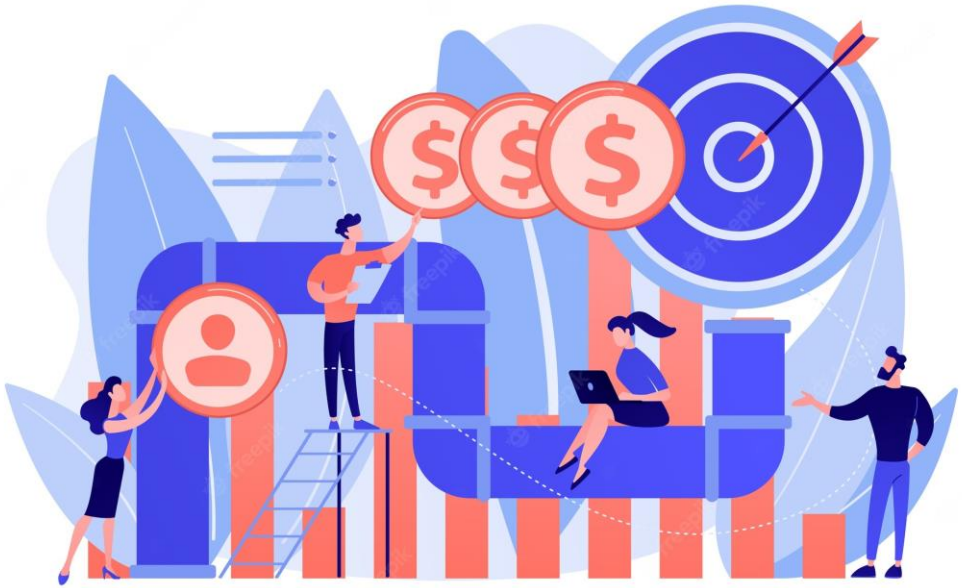
Analise de RFM por Cluster



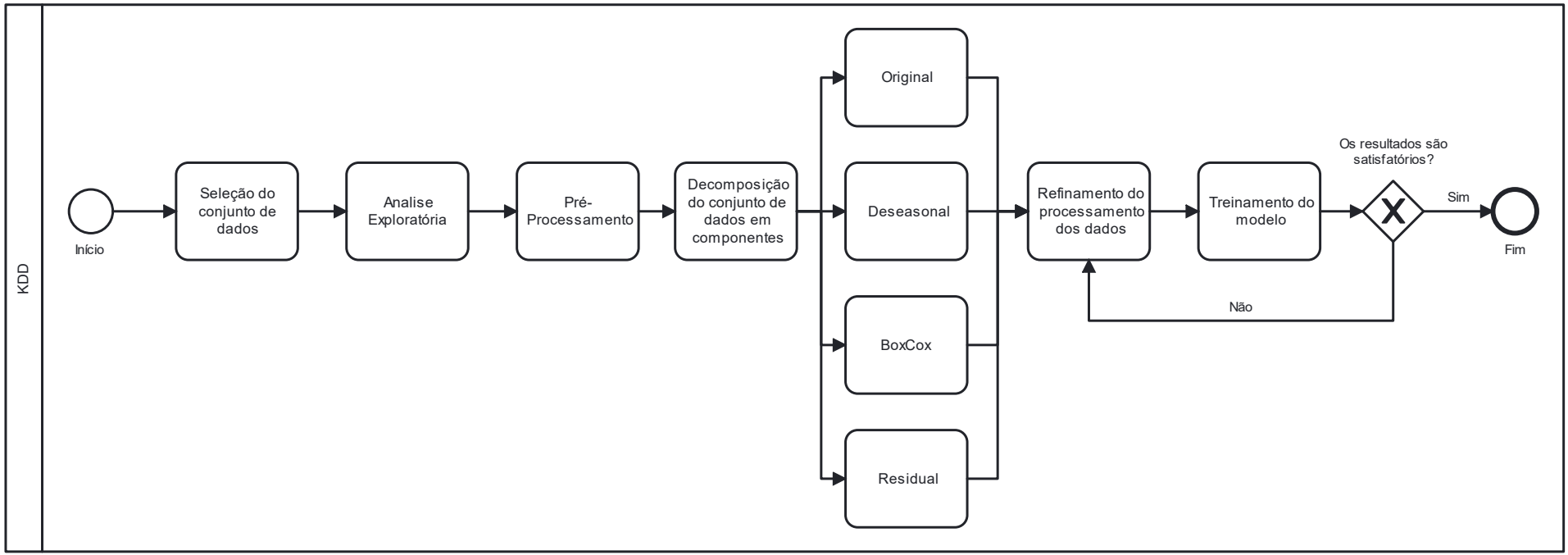




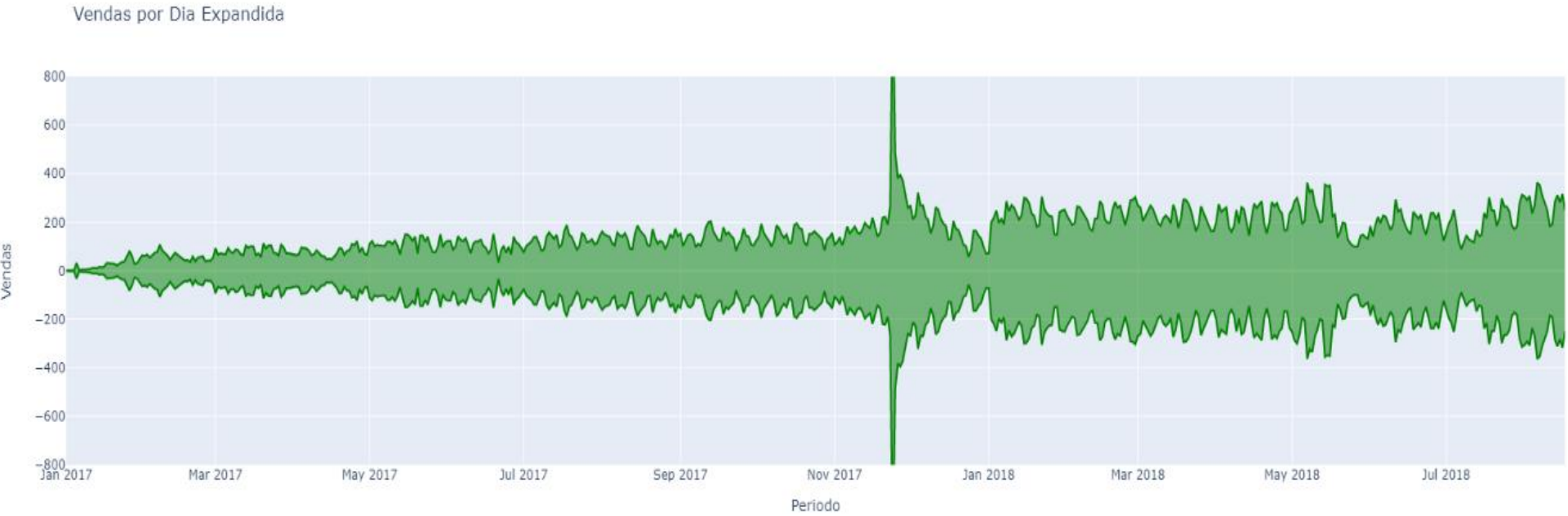
Séries Temporais

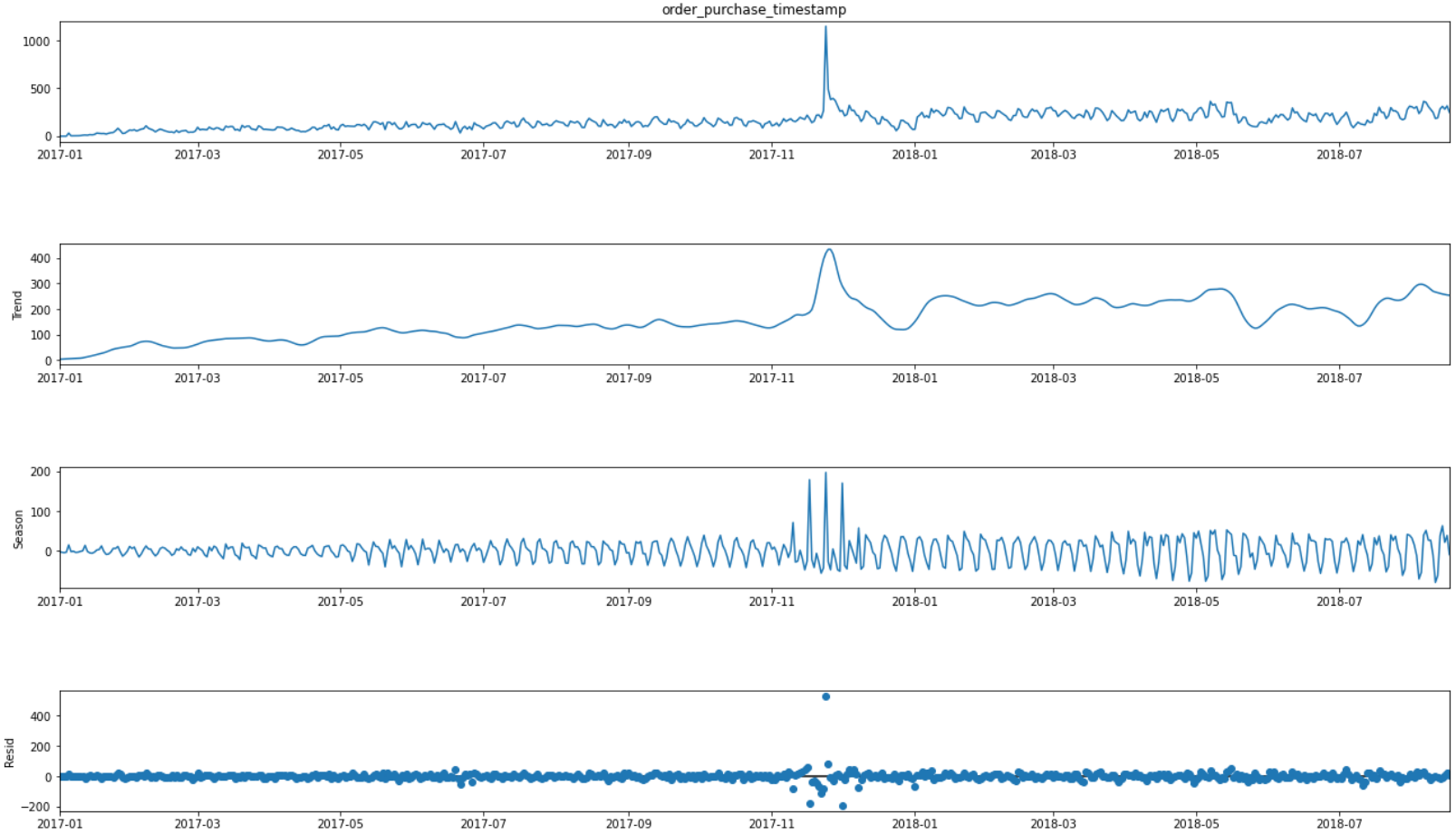


7. Séries Temporais



Vendas por dia desconsiderando os outliers



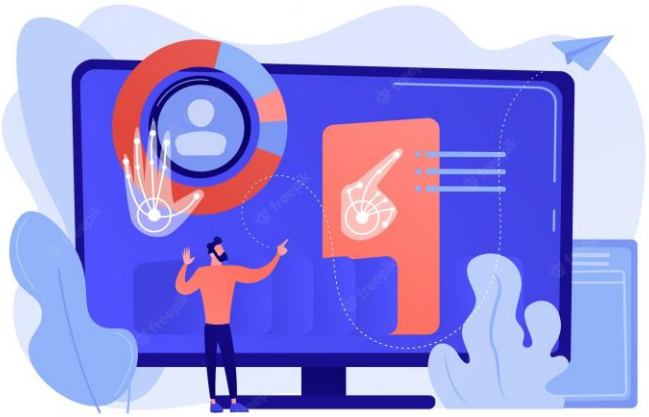

order_purchase_timestamp
1000
500
0
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
Trend
400
300
200
100
0
Season
200
100
0
Resid
400
200
0
−200

Conjunto de Dado	Descrição
Original	Dados originais sem nenhuma transformação.
Deseasonal	Dados em que se aplica a remoção da sazonalidade.
BoxCox	Dados que aplicam a transformação de Box Cox, com objetivo de aproximar os dados de forma ótima de uma normal.
Residual	Variações aleatórias na série temporal.

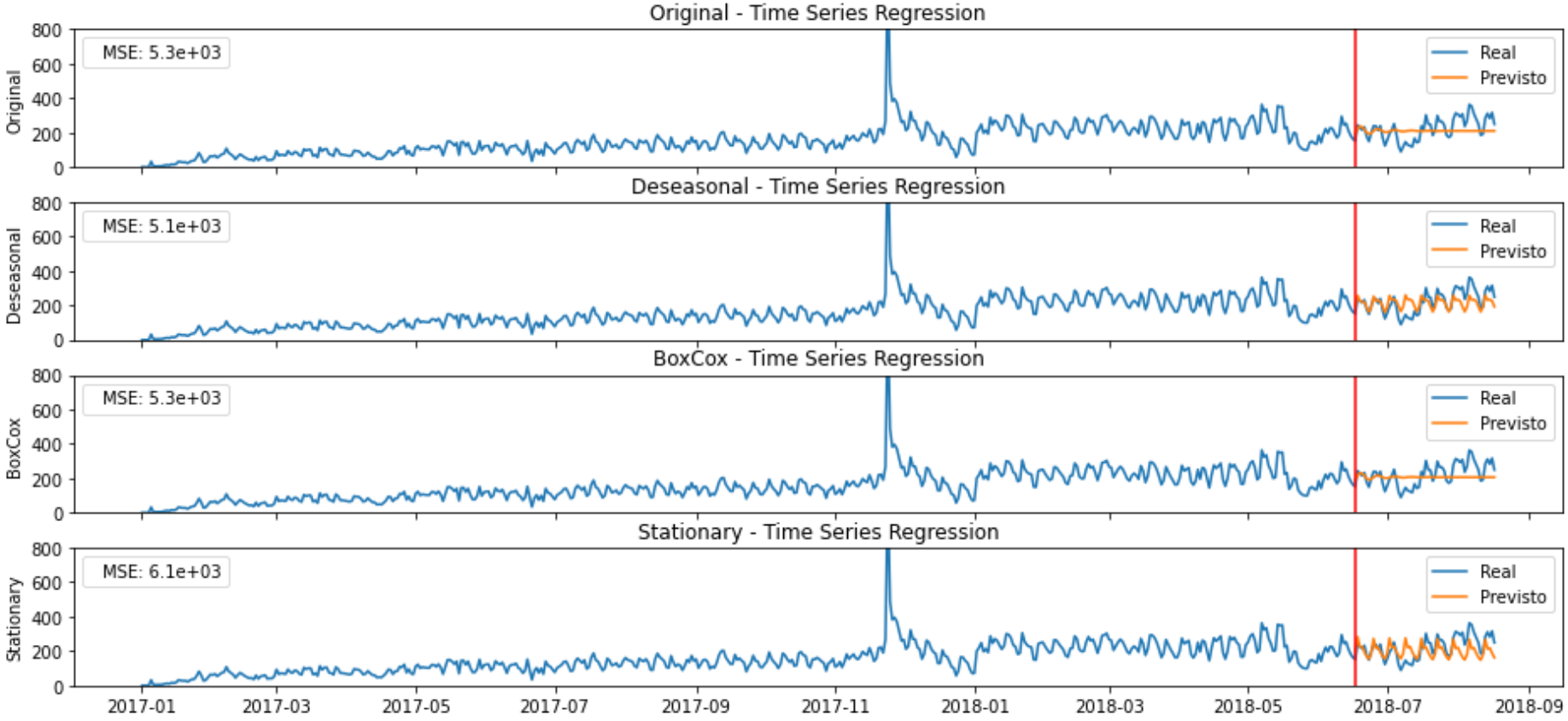
Grupo	Nível de Confiança	Estacionário
Original	99%	Não
Deseasonal	99%	Não
BoxCox	99%	Não
Residual	99%	Sim

7.1 Algoritmos

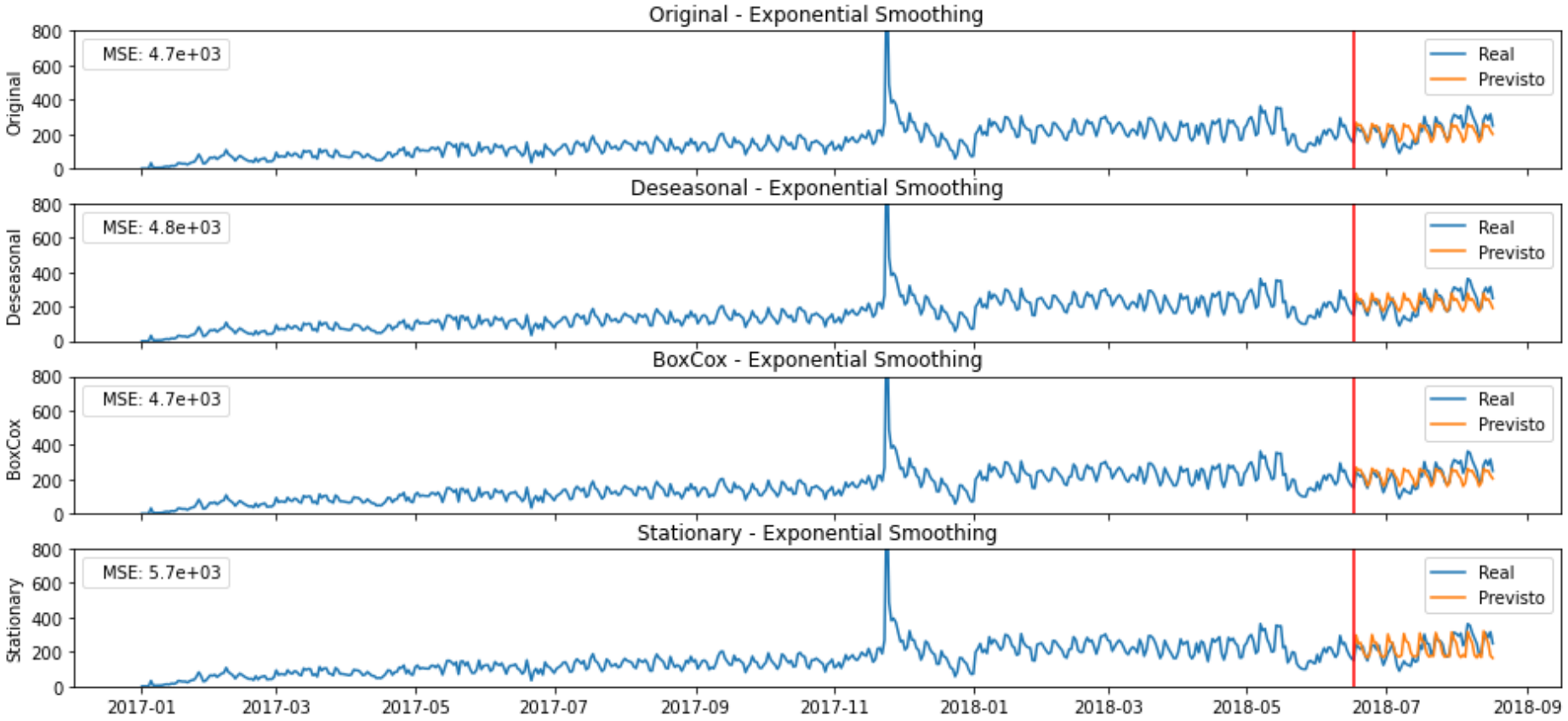
- Regressão Linear;
- Triple Exponential Smoothing (TES);
- Autoregressive integrated moving average (ARIMA);
- Long short-term memory (LSTM);
- XGBoost;



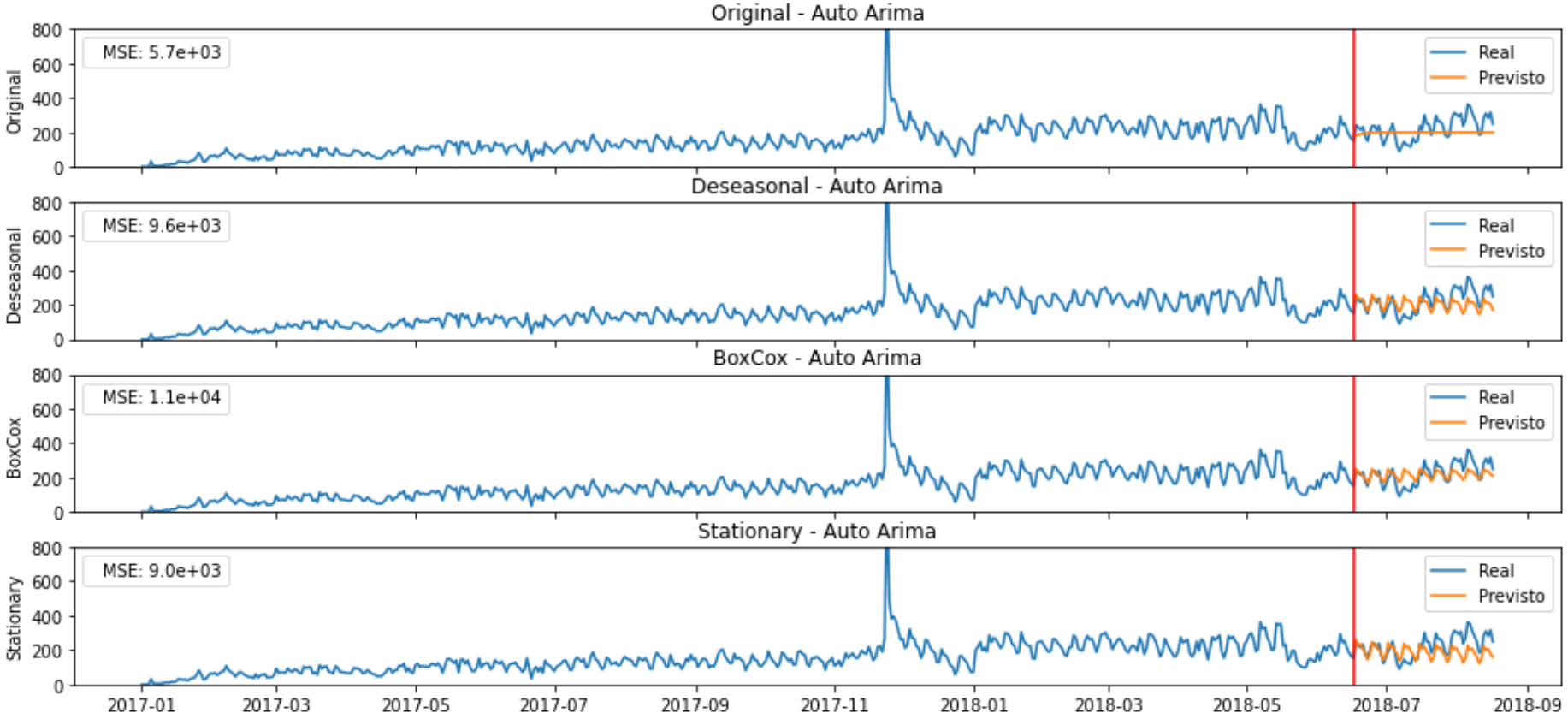
7.2 Regressão Linear



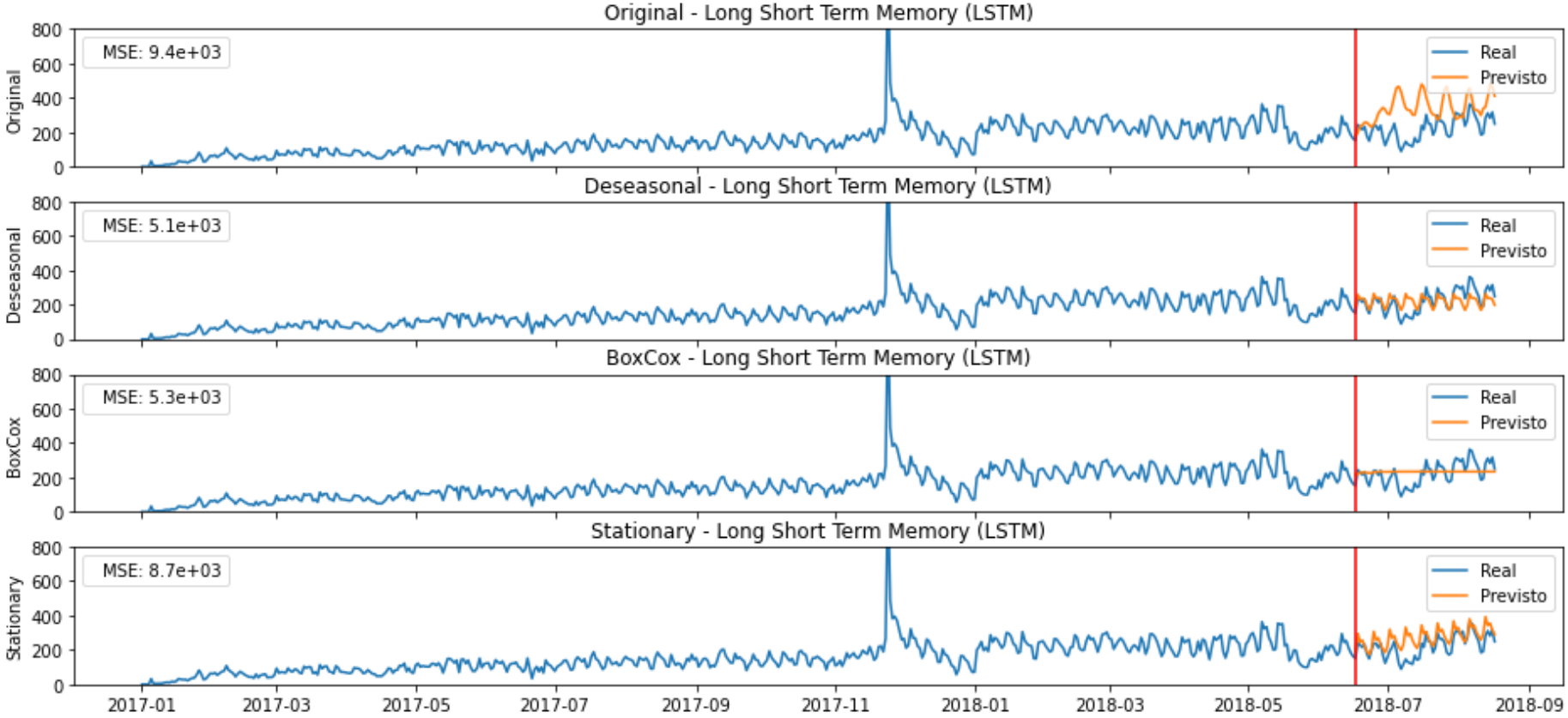
7.3 TES


Original - Exponential Smoothing
MSE: 4.7e+03
Original
Deseasonal - Exponential Smoothing
MSE: 4.8e+03
Deseasonal
BoxCox - Exponential Smoothing
MSE: 4.7e+03
BoxCox
Stationary - Exponential Smoothing
MSE: 5.7e+03
Stationary
Real
Previsto
0
200
400
600
800
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09

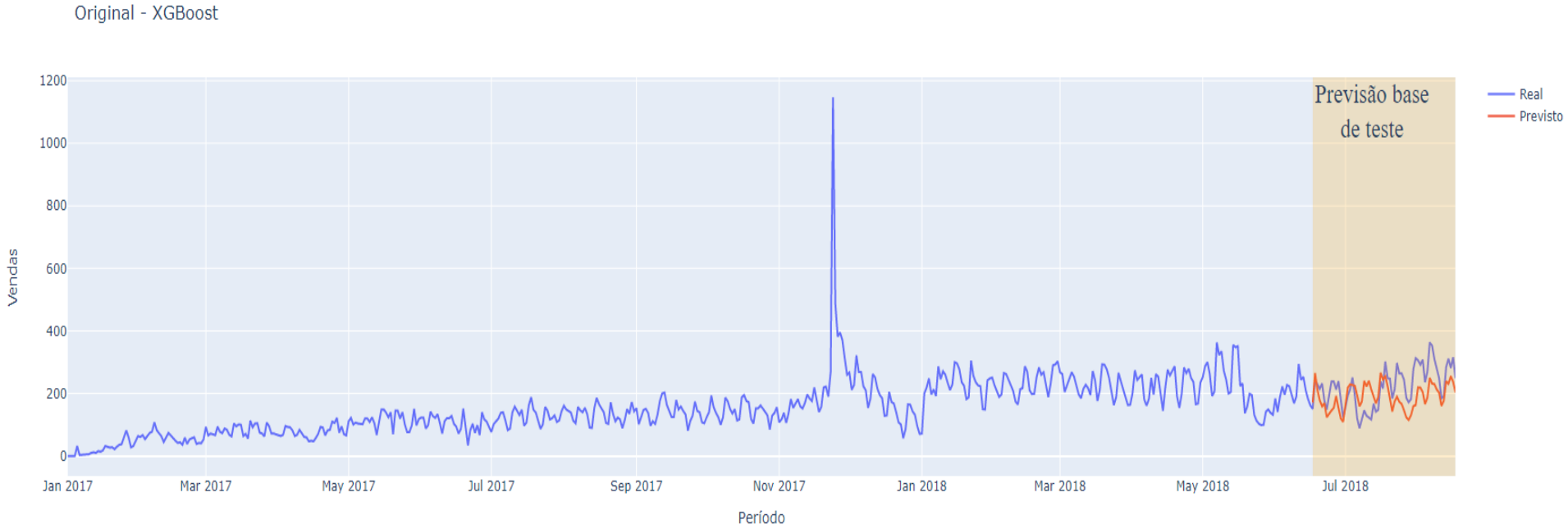
7.4 ARIMA


Original - Auto Arima
MSE: 5.7e+03
Original
Deseasonal - Auto Arima
MSE: 9.6e+03
Deseasonal
BoxCox - Auto Arima
MSE: 1.1e+04
BoxCox
Stationary - Auto Arima
MSE: 9.0e+03
Stationary
Real
Previsto
0
200
400
600
800
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09

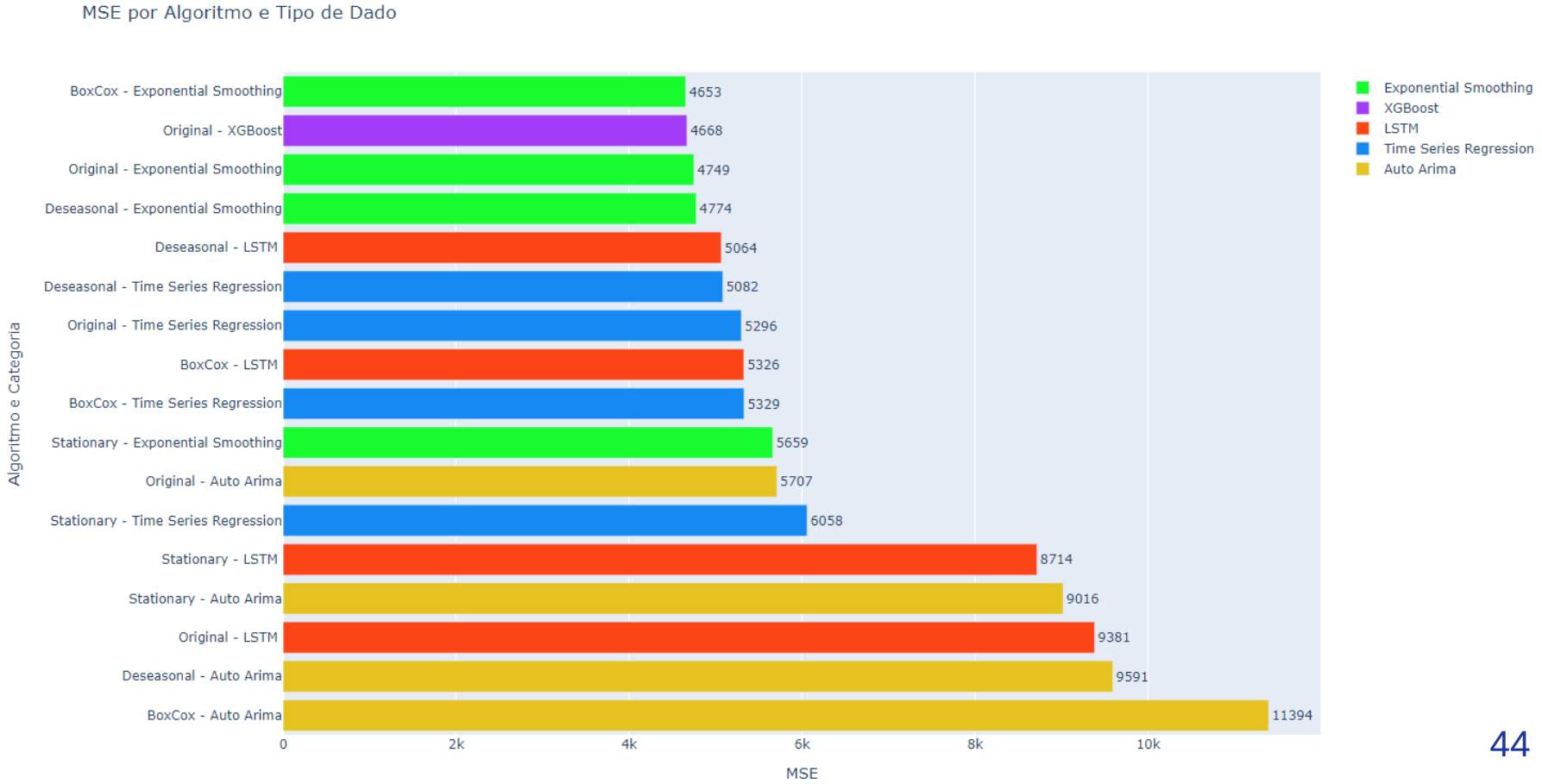
7.5 LSTM


Original - Long Short Term Memory (LSTM)
MSE: 9.4e+03
Original
Real
Previsto
Deseasonal - Long Short Term Memory (LSTM)
MSE: 5.1e+03
Deseasonal
BoxCox - Long Short Term Memory (LSTM)
MSE: 5.3e+03
BoxCox
Stationary - Long Short Term Memory (LSTM)
MSE: 8.7e+03
Stationary
0
200
400
600
800
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09

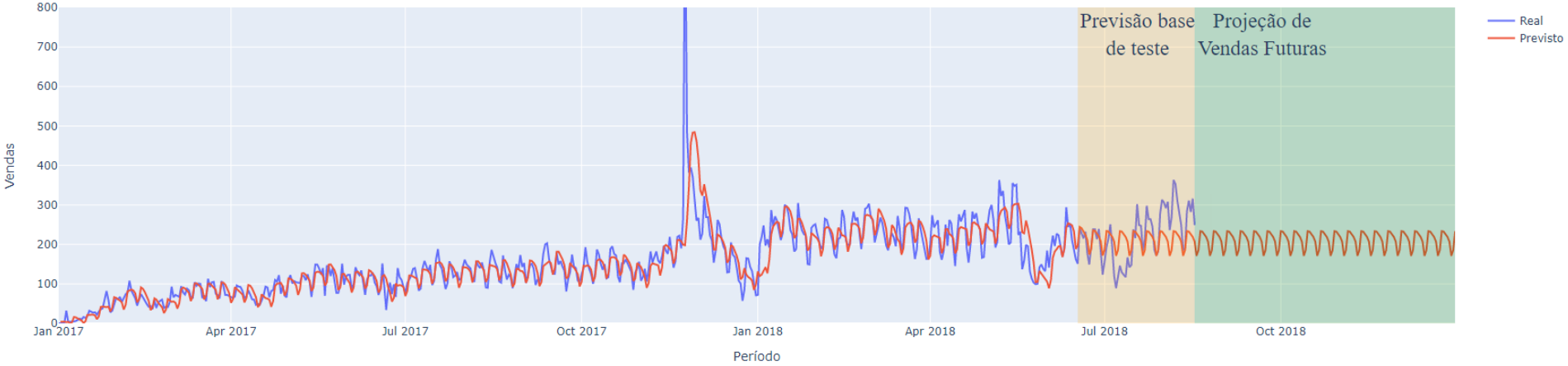
7.6 XGBoost


Original - XGBoost
1200
1000
800
600
400
200
0
Vendas
Jan 2017
Mar 2017
May 2017
Jul 2017
Sep 2017
Nov 2017
Jan 2018
Mar 2018
May 2018
Jul 2018
Período
Previsão base de teste
Real
Previsto

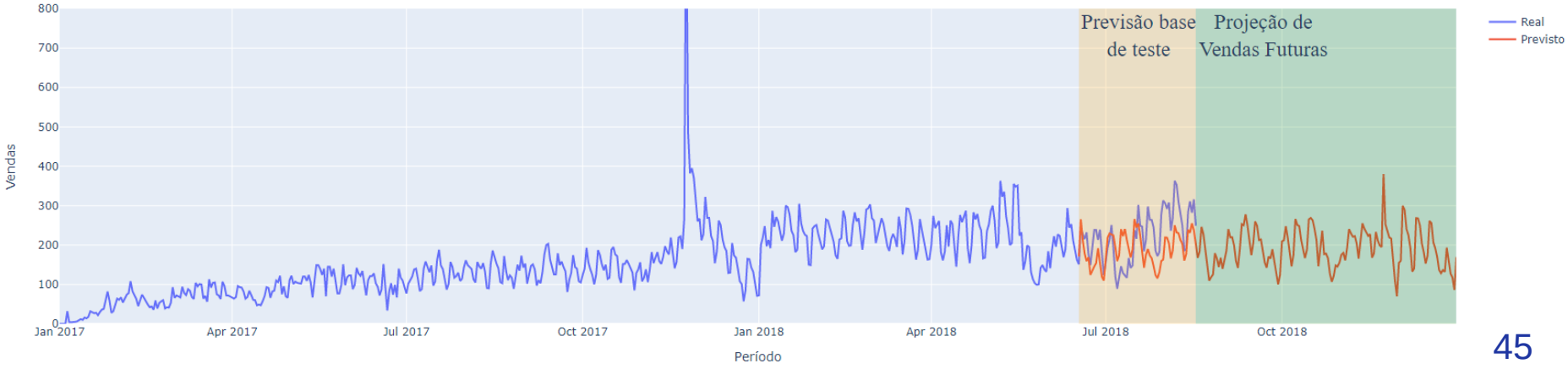
7.7 Compaação Modelos



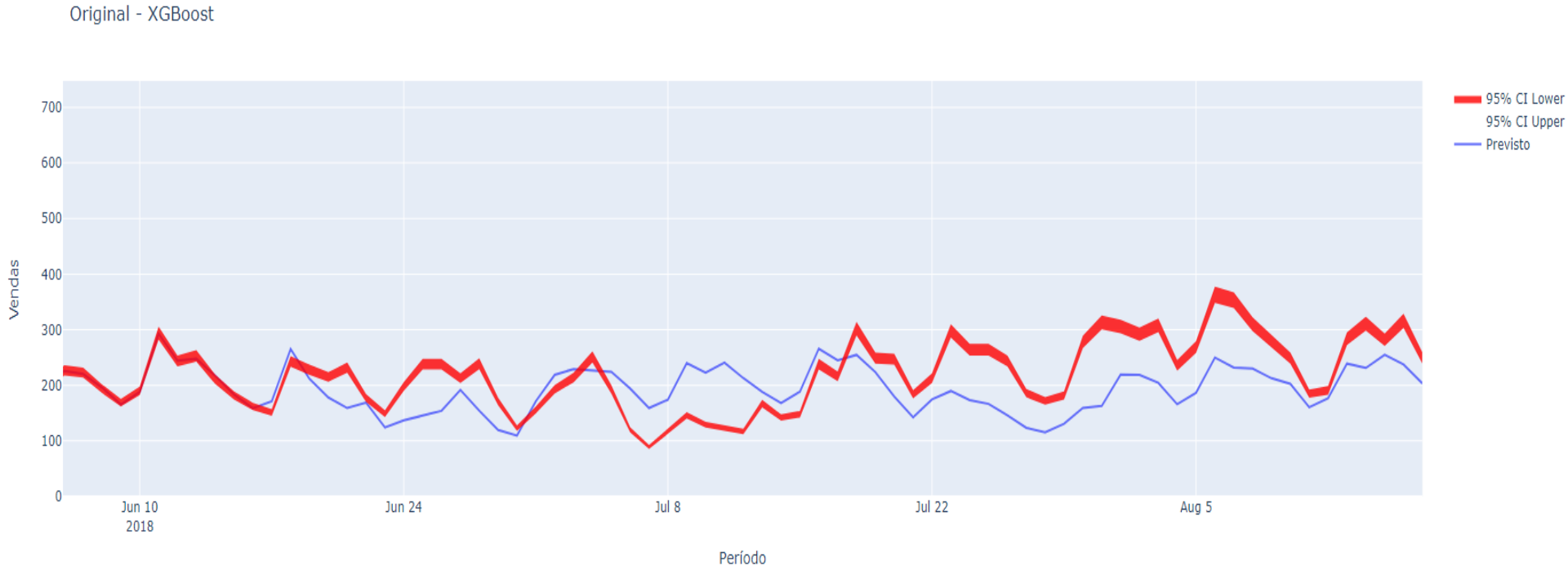
BoxCox - Exponential Smoothing



Original - XGBoost

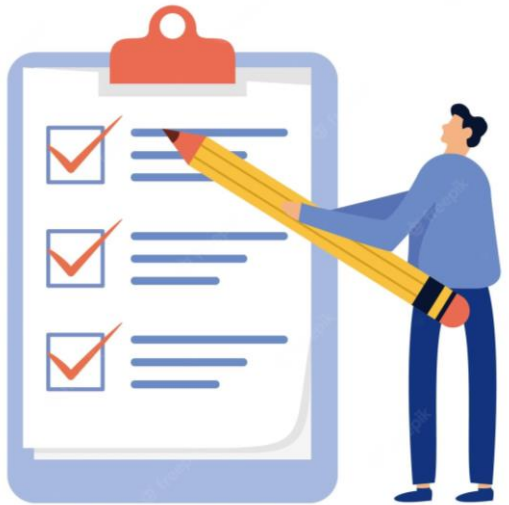


Limites de confiança de 95% para os valores reais.



8. Conclusão

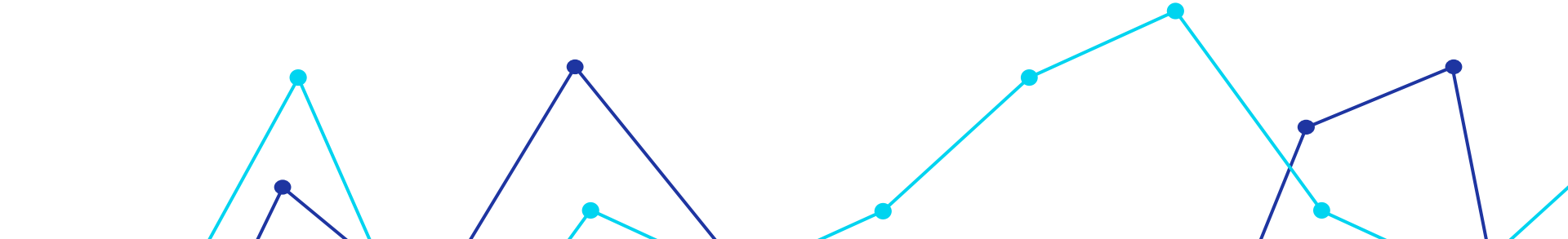
- Definição de atributos mais relevantes;
- Definição de algoritmos de Mineração de Dados;
- Aplicação de técnicas de Mineração de Dados;
- Avaliação de resultados obtidos.



9. Trabalhos Futuros

- Continuação do estudo do comércio eletrônico, abordando outros algoritmos, fontes de dados e métricas;
- Ampliação do espaço de hipótese e parâmetros testados;
- Busca de soluções para manter o consumidor mais ativo no comércio.

OBRIGADO



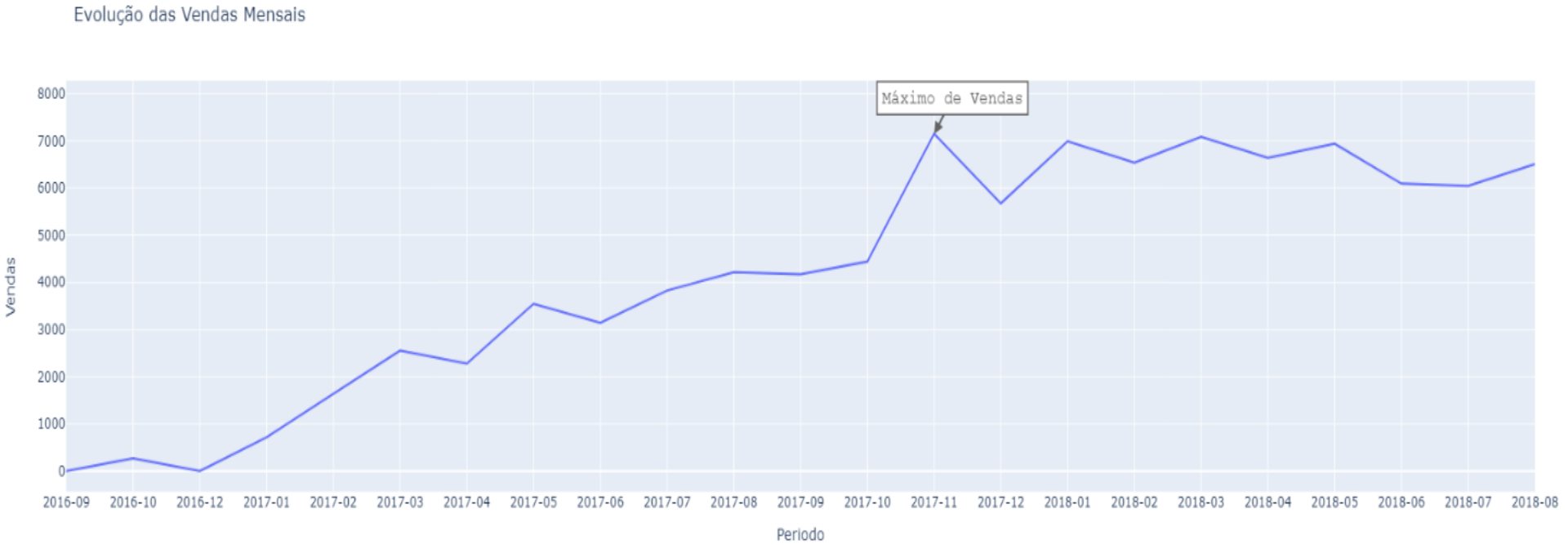
Metodologia

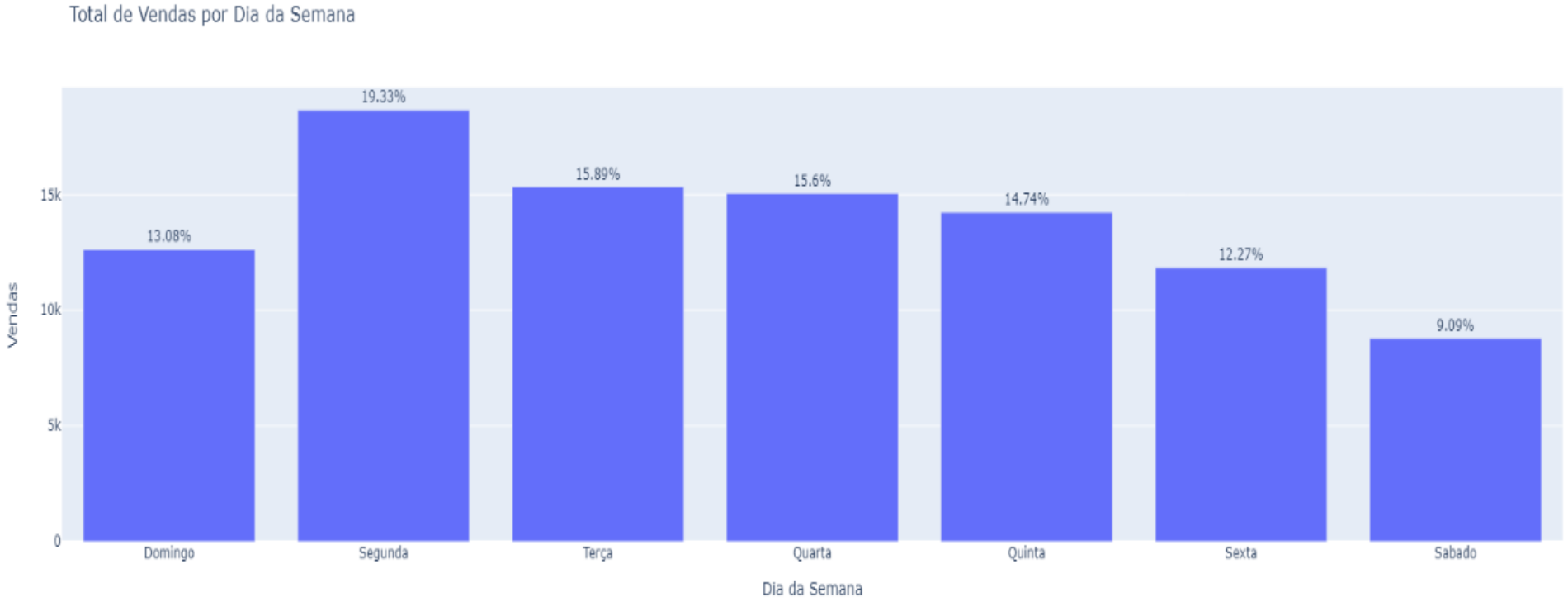
O presente trabalho utiliza o método indutivo. Tratando-se de uma pesquisa aplicada, realizada com objetivo de obter novos conhecimentos para responder as perguntas de pesquisa do projeto.

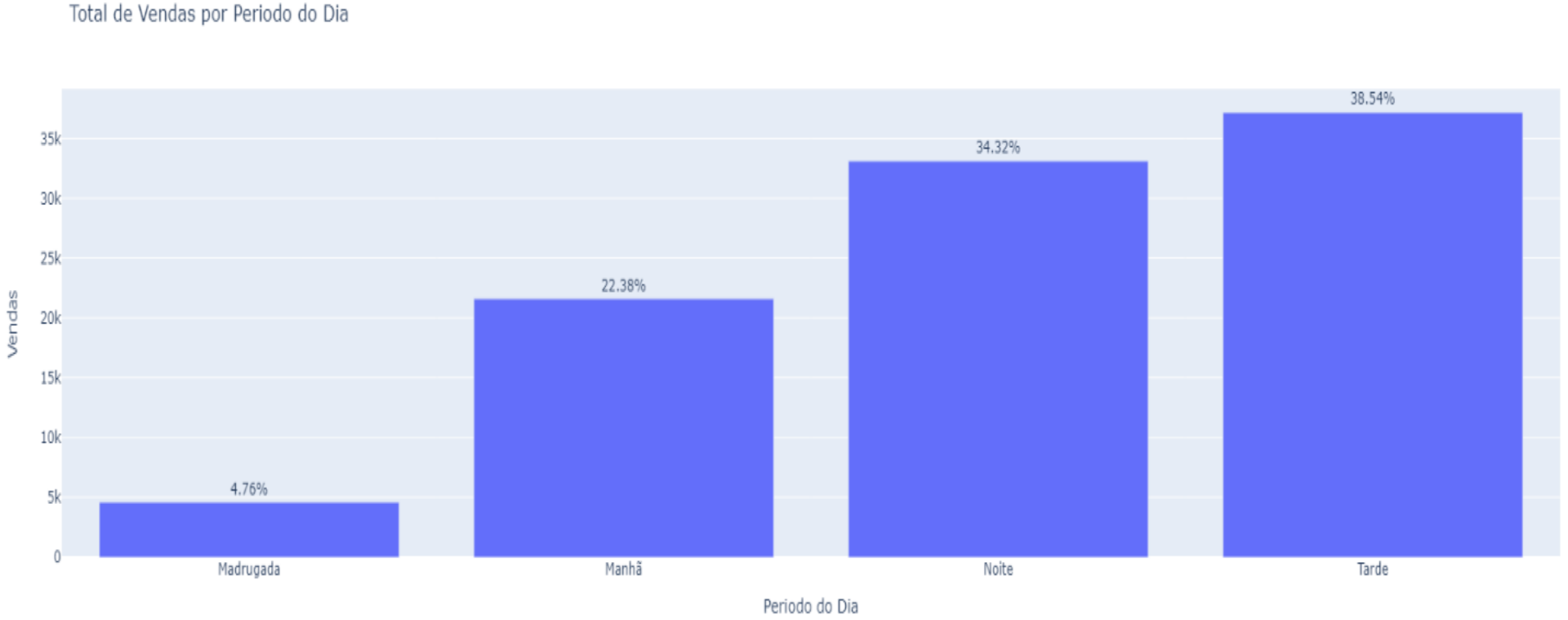
4. Trabalhos Relacionados

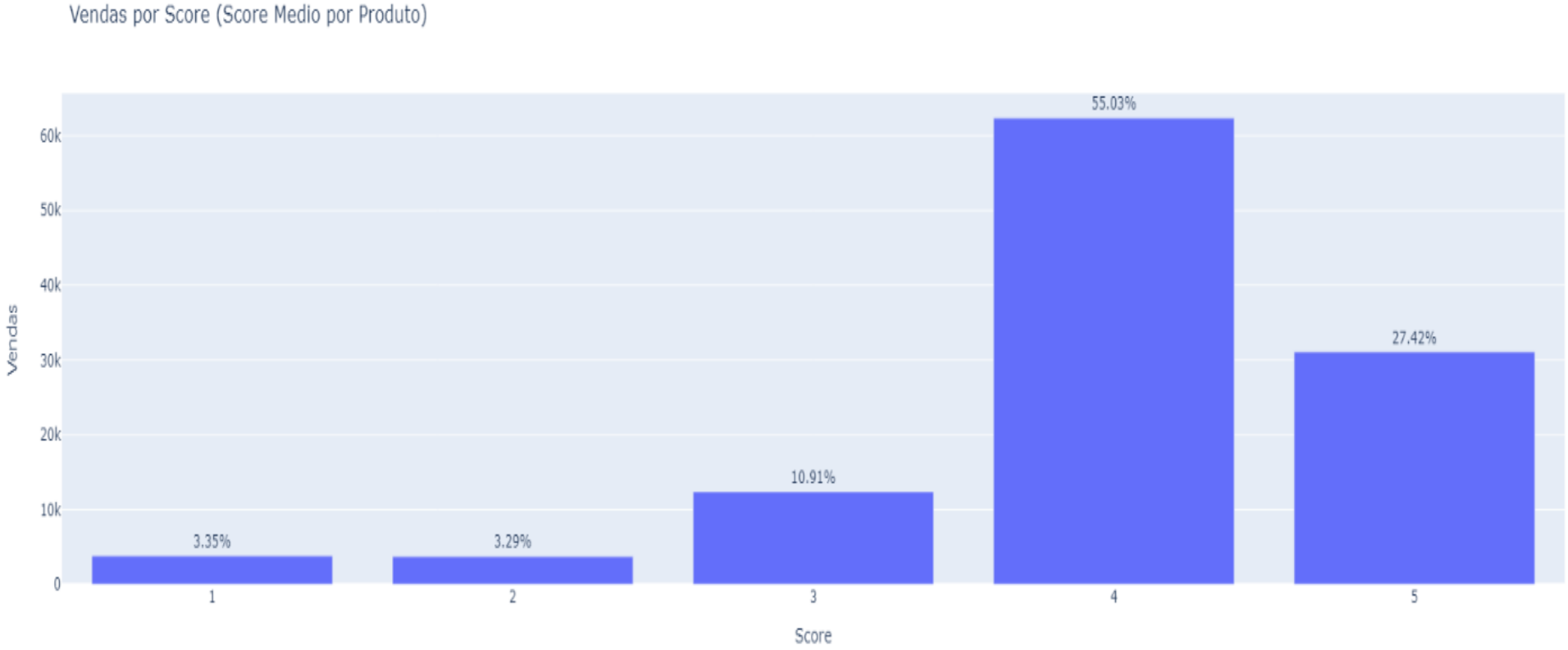
ID	Autores	Ano	Título	Fonte de dados
1	Seonghoon Moon, Suman Bae, Songkuk Kim	2014	Predicting the Near-Weekend Ticket Sales Using Web-Based External Factors and Box-Office Data	ACM Digital Libraty
2	Patcharin Ponyiam, Somjit Arch-int	2018	Customer Behavior Analysis Using Data Mining Techniques	IEEEXplore
3	Guixiang Zhua, Zhiang Wub, Youquan Wangb, Shanshan Caob, Jie Cao	2019	Online purchase decisions for tourism e-commerce	ScienceDirect

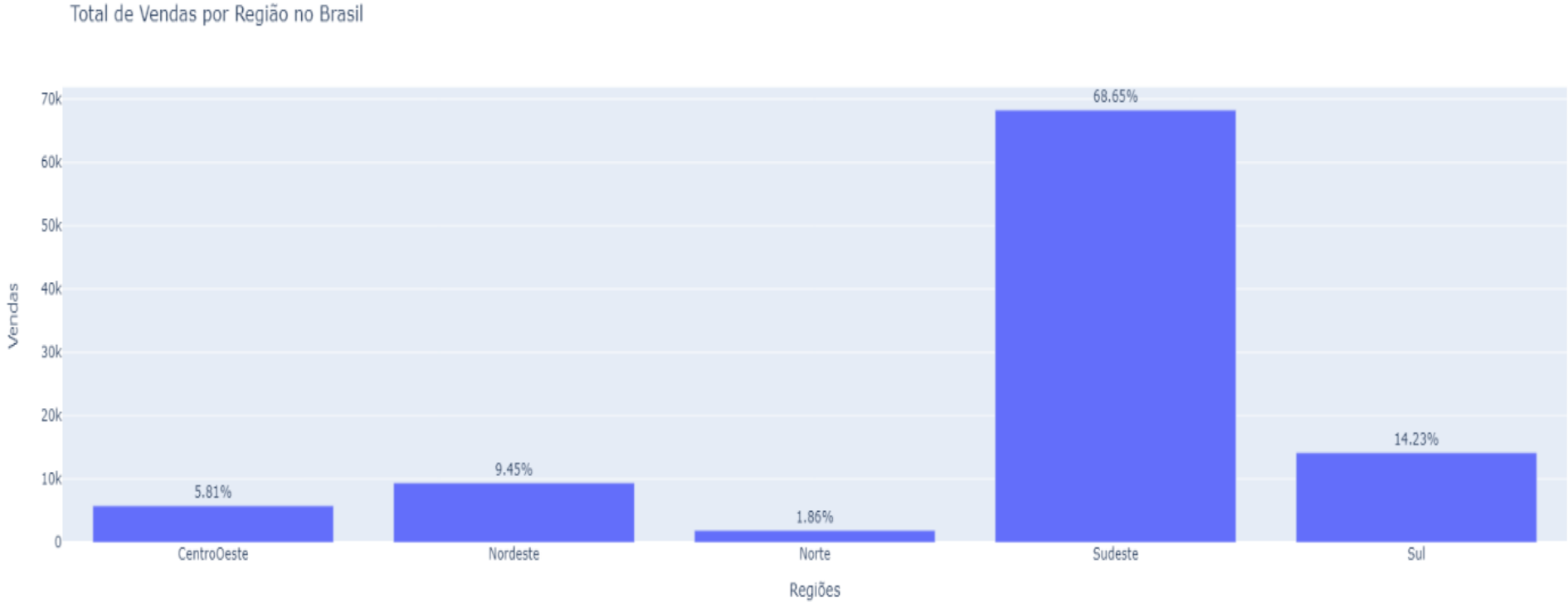
O período com maior quantidade de vendas foi Novembro de 2017.

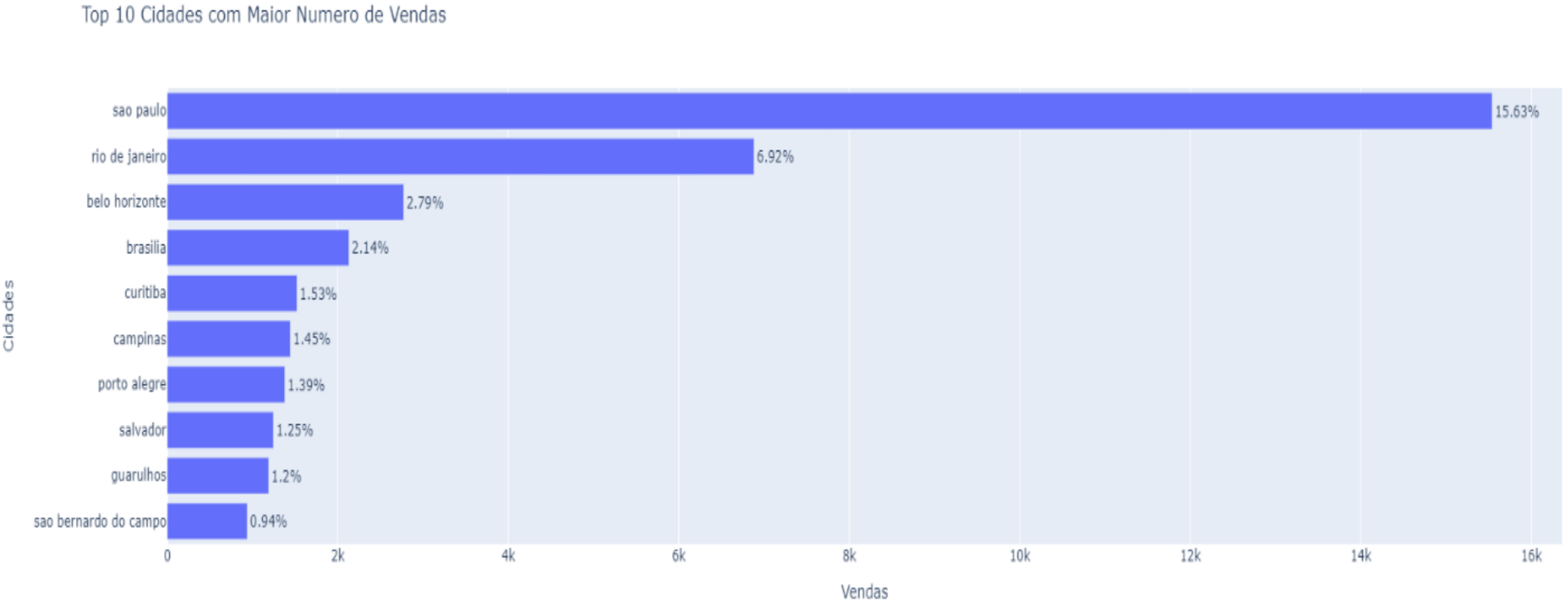


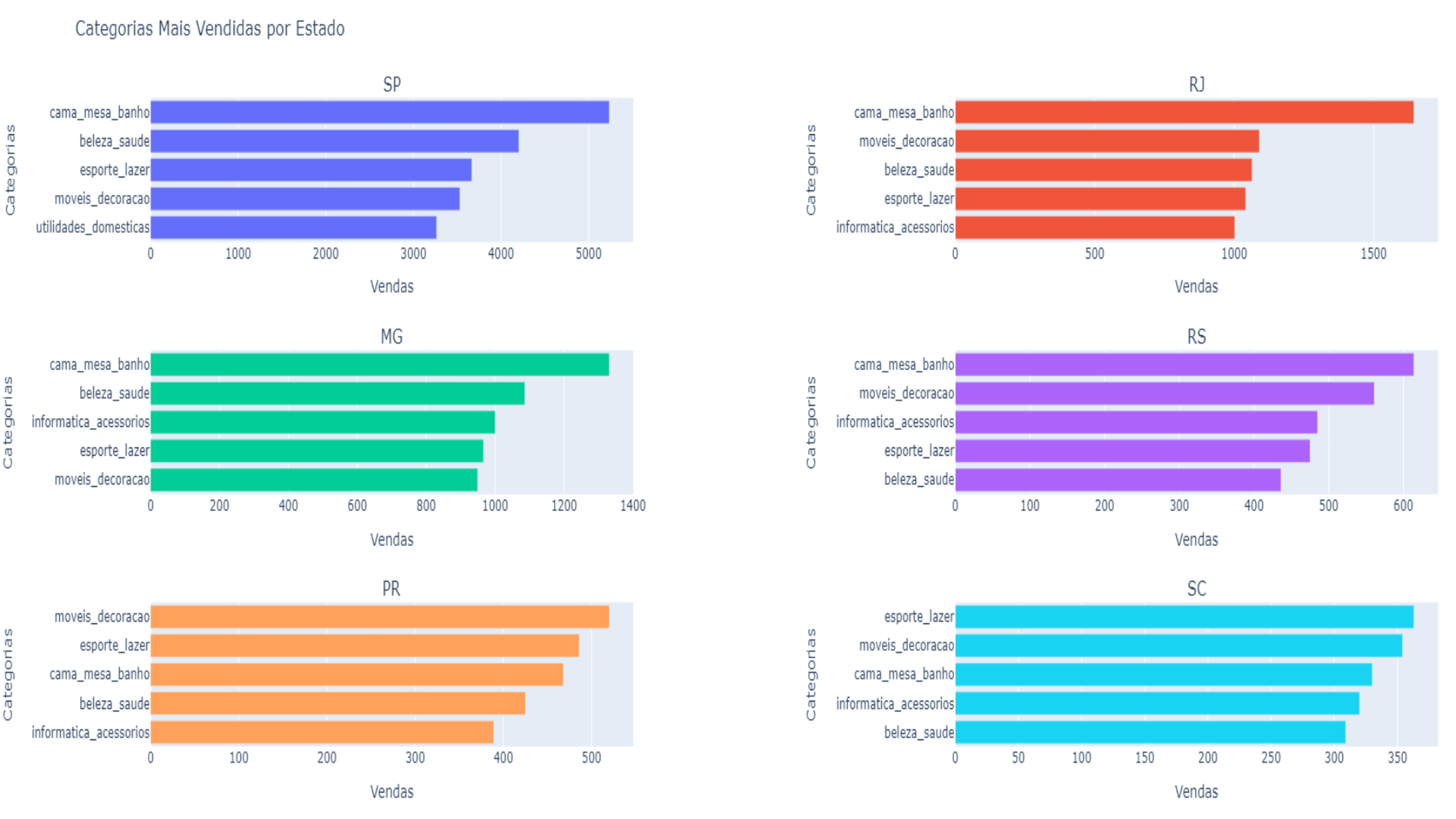

Total de Vendas por Dia da Semana
19.33%
15.89%
15.6%
14.74%
13.08%
12.27%
9.09%
15k
10k
5k
0
Vendas
Domingo
Segunda
Terça
Quarta
Quinta
Sexta
Sabado
Dia da Semana


Total de Vendas por Periodo do Dia
Vendas
35k
30k
25k
20k
15k
10k
5k
0
4.76%
22.38%
34.32%
38.54%
Madrugada
Manhã
Noite
Tarde
Periodo do Dia

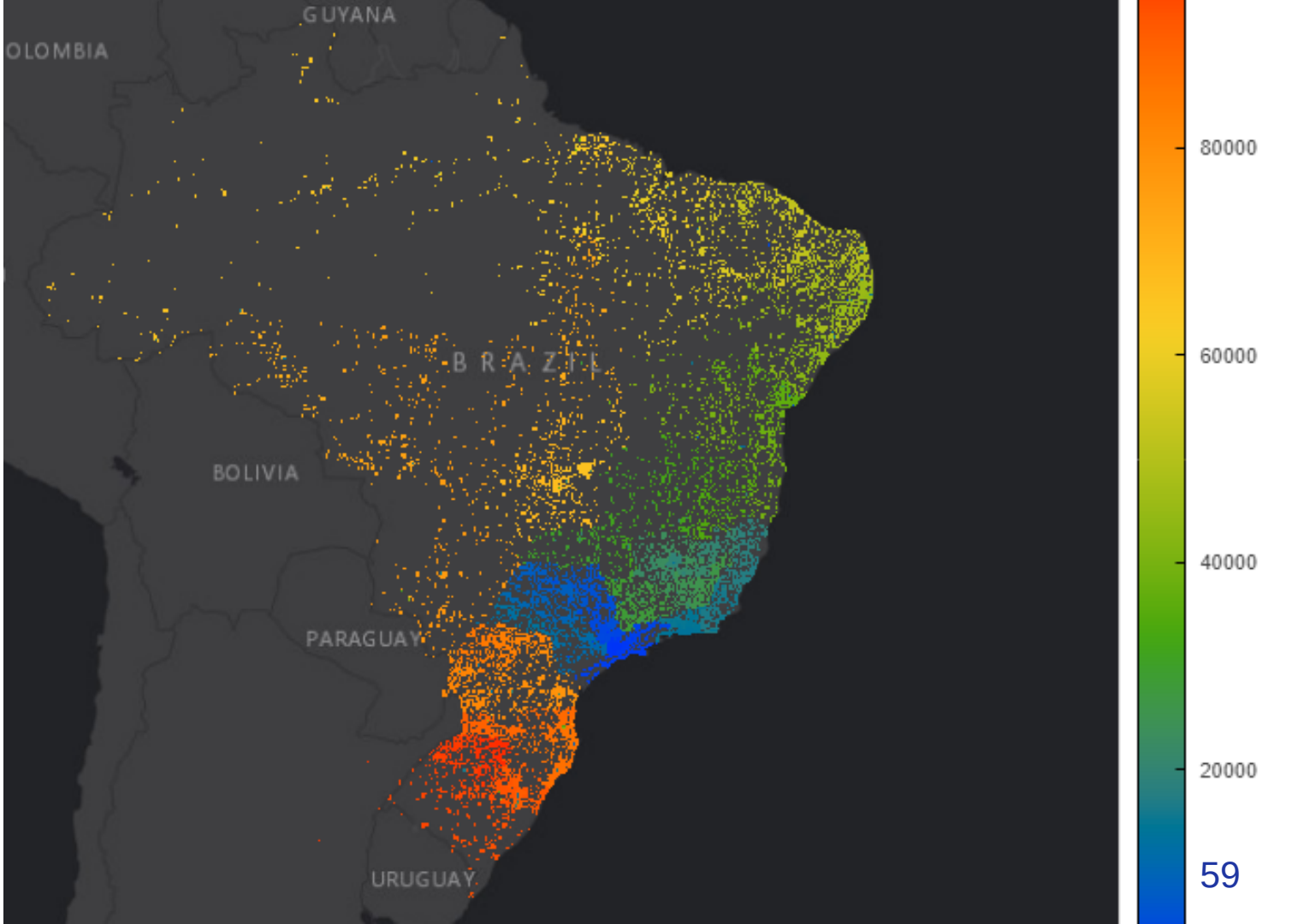

Vendas por Score (Score Medio por Produto)
Vendas
60k
50k
40k
30k
20k
10k
0
3.35%
3.29%
10.91%
55.03%
27.42%
1
2
3
4
5
Score


Total de Vendas por Região no Brasil
70k
60k
50k
40k
30k
20k
10k
0
Vendas
5.81%
9.45%
1.86%
68.65%
14.23%
CentroOeste
Nordeste
Norte
Sudeste
Sul
Regiões

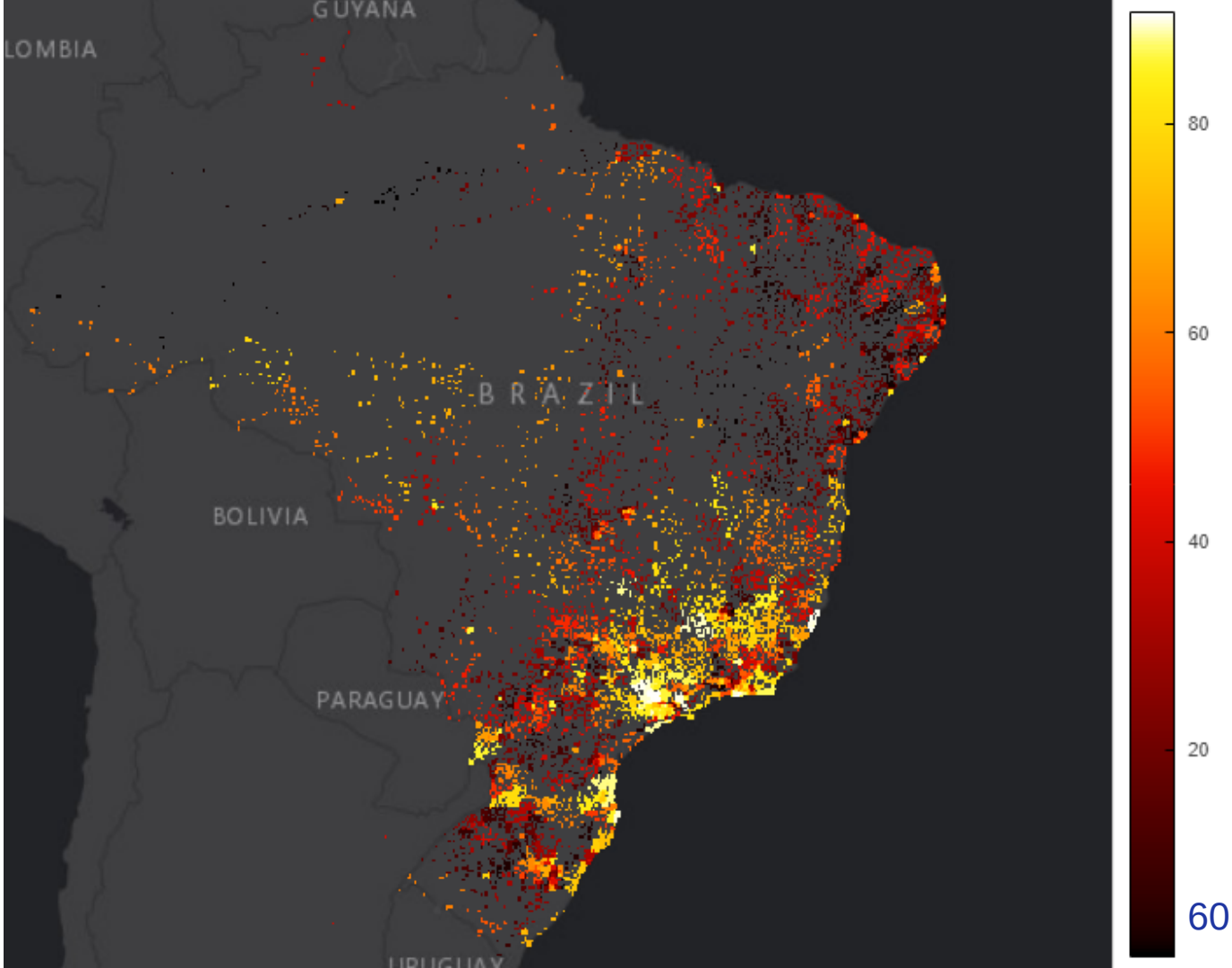

Top 10 Cidades com Maior Numero de Vendas
sao paulo
rio de janeiro
belo horizonte
brasilia
curitiba
campinas
porto alegre
salvador
guarulhos
sao bernardo do campo
15.63%
6.92%
2.79%
2.14%
1.53%
1.45%
1.39%
1.25%
1.2%
0.94%
Cidades
0
2k
4k
6k
8k
10k
12k
14k
16k
Vendas


Categorias Mais Vendidas por Estado
SP
cama_mesa_banho
beleza_saude
esporte_lazer
moveis_decoracao
utilidades_domesticas
Categorias
0
1000
2000
3000
4000
5000
Vendas
RJ
cama_mesa_banho
moveis_decoracao
beleza_saude
esporte_lazer
informatica_acessorios
Categorias
0
500
1000
1500
Vendas
MG
cama_mesa_banho
beleza_saude
informatica_acessorios
esporte_lazer
moveis_decoracao
Categorias
0
200
400
600
800
1000
1200
1400
Vendas
RS
cama_mesa_banho
moveis_decoracao
informatica_acessorios
esporte_lazer
beleza_saude
Categorias
0
100
200
300
400
500
600
Vendas
PR
moveis_decoracao
esporte_lazer
cama_mesa_banho
beleza_saude
informatica_acessorios
Categorias
0
100
200
300
400
500
Vendas
SC
esporte_lazer
moveis_decoracao
cama_mesa_banho
informatica_acessorios
beleza_saude
Categorias
0
50
100
150
200
250
300
350
Vendas

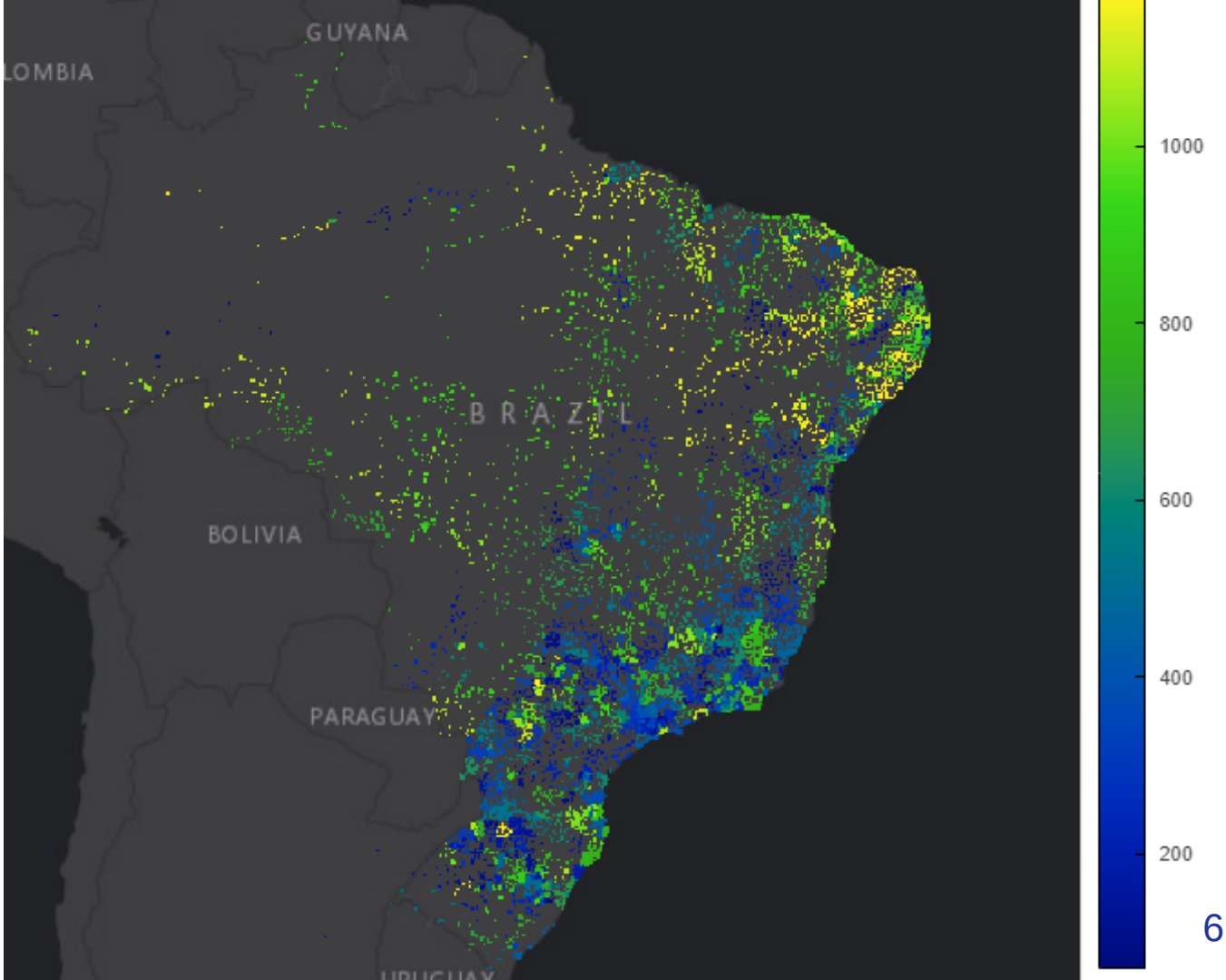
Pedidos por CEP do Brasil



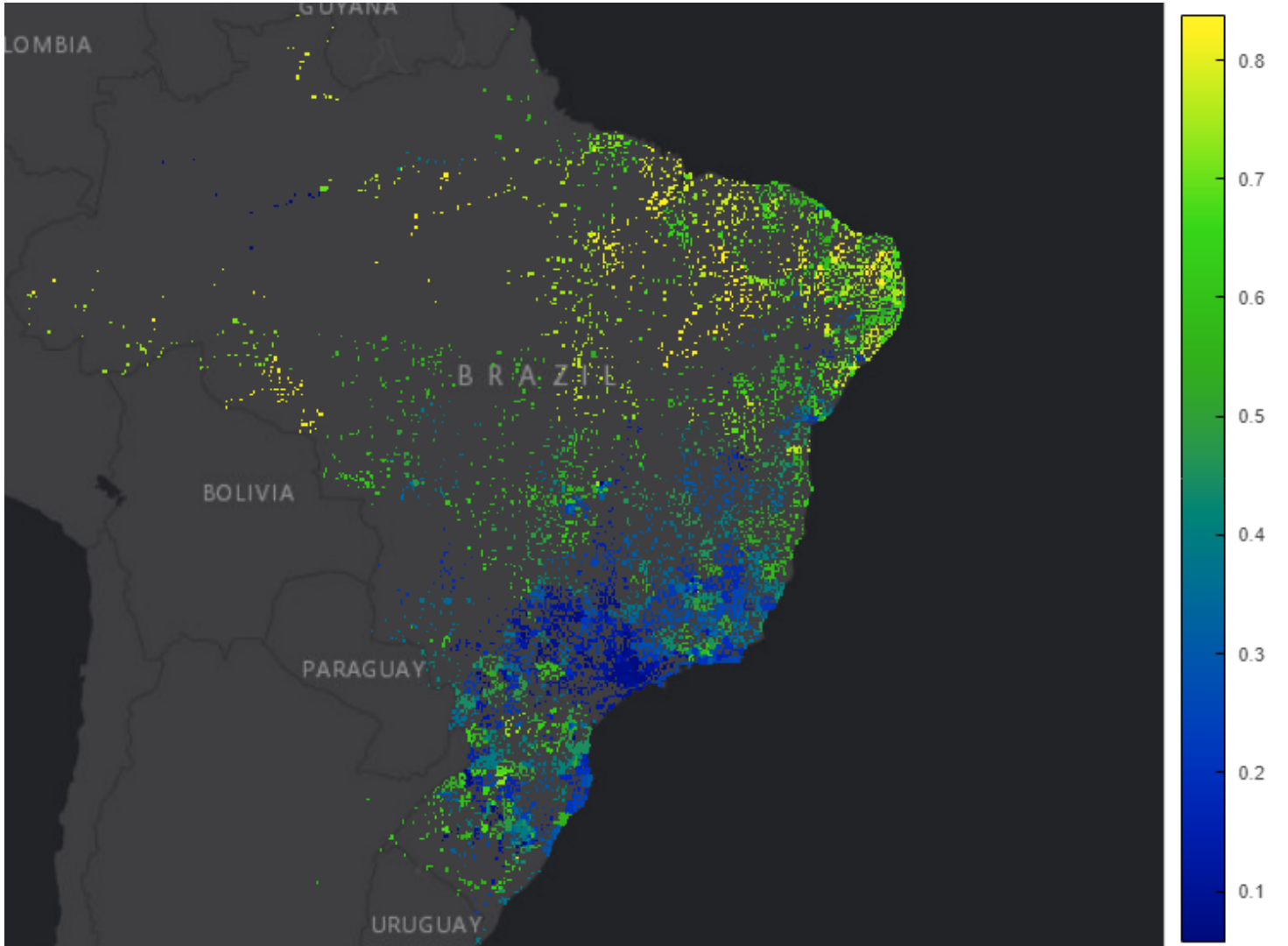
Faturamento por região no Brasil



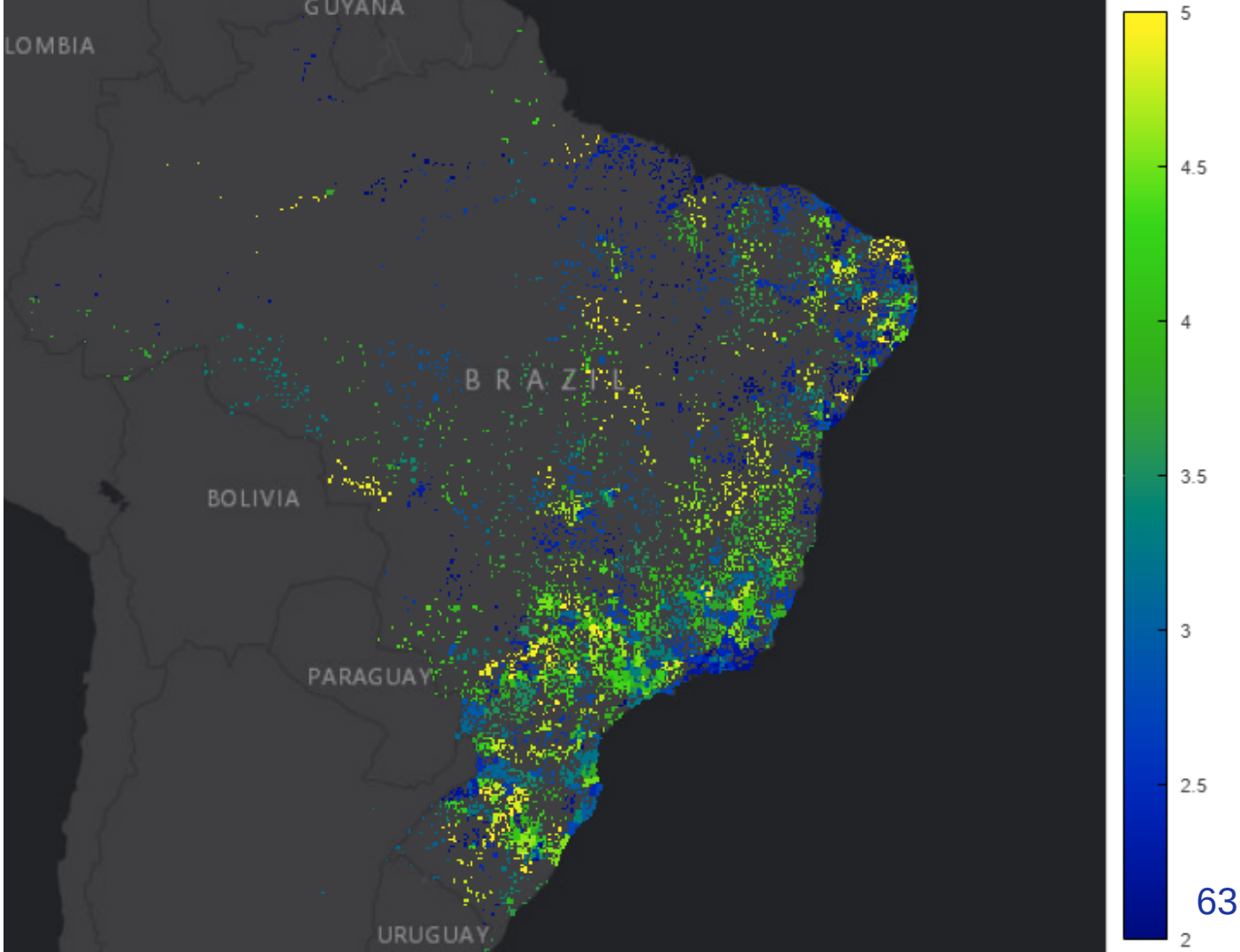
Ticket médio por região no Brasil



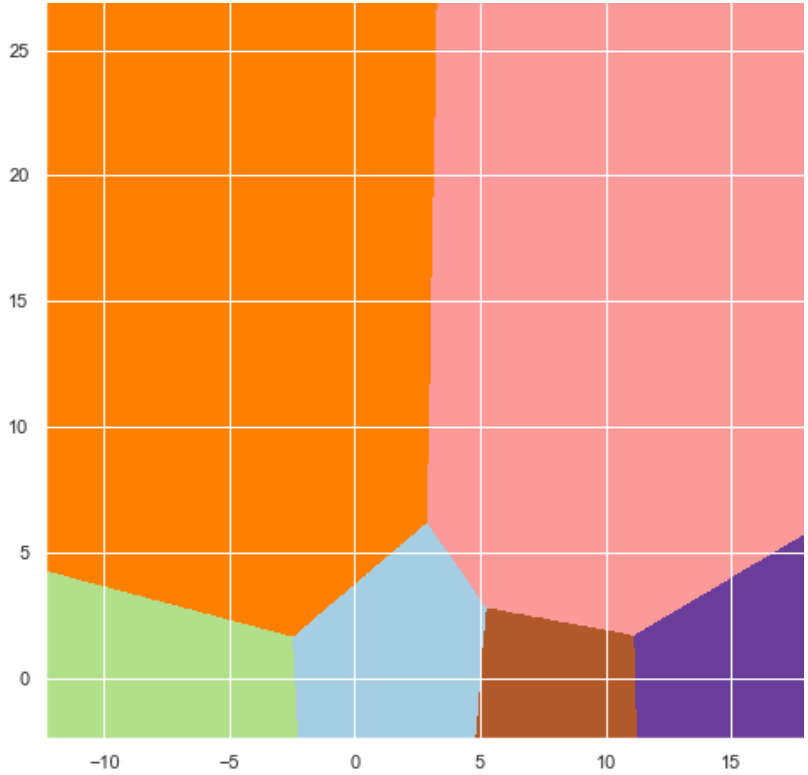
Média de Frete por Região no Brasil



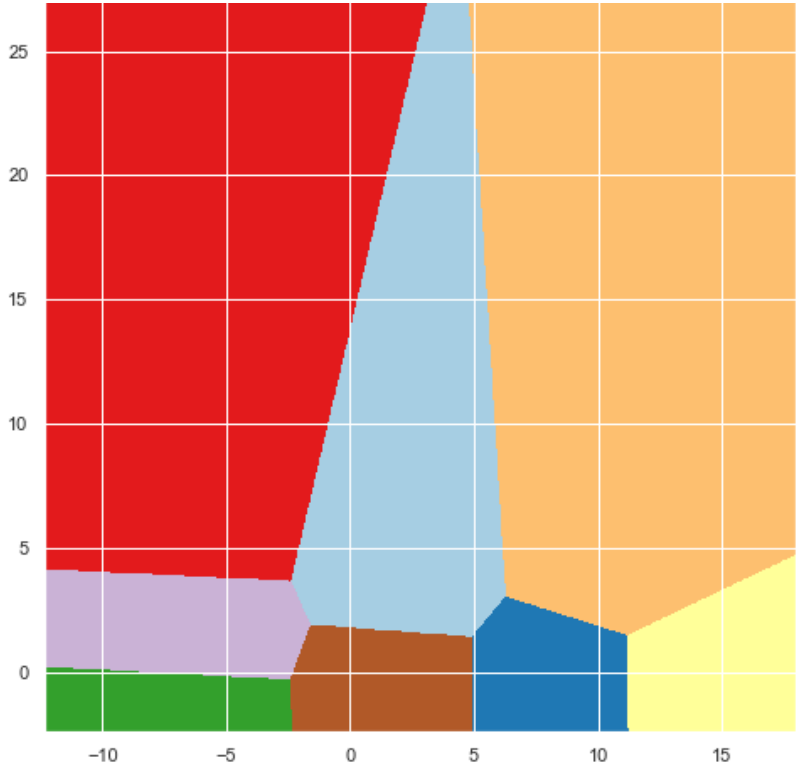
Review Médio por Região no Brasil



Meshgrid de 6 clusters com PCA



Meshgrid de 8 clusters com PCA



Clusterização Simples

Cluster	Preço	Frete	Parcelas	Score	Itens	RFM	Pagamento	Estado	Região
0	165.98	23.13	2.4	3.96	1.33	2.95	Crédito	RJ	Sudeste
1	130.2	15.62	1.77	4.16	1.37	3.05	Crédito	SP	Sudeste
2	187.74	23.64	3.16	4.01	1.38	2.97	Crédito	MG	Sudeste
3	320.17	23.27	8.41	3.98	1.67	3.03	Crédito	SP	Sudeste

Clusterização por RFM

Cluster	Recency	Frequency	Monetary	RFM
0	395.67	1	296.31	1.64
1	133.88	1	296.32	3.9
2	221	2.23	682.74	3.27
3	221.85	1.49	26987.42	3.2

6.4 LSTM

Atributo	Valor / Parâmetro
Camada de entrada LSTM	Input_shape: (7, 1)
Camada Densa com 1 neurônio de entrada	Initializer: GlorotUniform
Camada de BatchNormalization	N/A
Camada Densa com 1 neurônio de entrada	Initializer: GlorotUniform
EarlyStopping	Monitor: loss Patience: 3 Mode: min
Otimizador Adam	Taxas de aprendizado testadas: 0.00001; 0.0001; 0.001; 00.1
Função de perda	MSE
Épocas	1000
StandarScaler	Pré-treinamento