UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

João Luiz Pacher

Previsão de Vendas com modelos de Aprendizado de Máquina

São Carlos

2022

João Luiz Pacher

Previsão de Vendas com modelos de Aprendizado de Máquina

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

São Carlos

2022

João Luiz Pacher

Previsão de Vendas com modelos de Aprendizado de Máquina

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Data de defesa: 5 de março de 2022

Comissão Julgadora:

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
Orientador

Professor
Convidado1

Professor
Convidado2

São Carlos
2022

RESUMO

PACHER, J. L. **Previsão de Vendas com modelos de Aprendizado de Máquina**. 2022. 74p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

O objeto desse trabalho foi estudar diferentes modelos de aprendizado de máquina para a tarefa de regressão com séries temporais.

A aplicação escolhida foi a previsão de vendas que é provavelmente o processo mais importante do gerenciamento da cadeia de *Supply Chain* das empresas e suporte aos negócios.

Foram utilizados desde modelos de grande capacidade até os mais simples: Redes Neurais, SARIMA, Holt-Winters e Theta, bem como modelos *Naive* que serviram de *baseline*.

Foram utilizadas medidas de desempenho comumente usadas em estatística e ciência de dados bem como outras específicas de séries temporais.

Palavras-chave: Previsão de Vendas. Séries Temporais. Redes Neurais. LSTM. SARIMA. Holt-Winters. Theta.

ABSTRACT

PACHER, J. L. **Sales Forecasting with Machine Learning models**. 2022. 74p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

The object of this work was to study different machine learning models for the time series regression task.

The application chosen was sales forecasting, which is probably the most important process of Supply Chain management of companies and support to the businesses.

From high-capacity to simple models were used: Neural Networks, SARIMA, Holt-Winters and Theta, as well as Naive models that served as baseline.

Performance metrics commonly used in statistics and data science and the ones specific for time series were used.

Keywords: Sales Forecasting. Time Series. Neural Networks. LSTM. SARIMA. Holt-Winters. Theta.

LISTA DE FIGURAS

Figura 1 – Avaliação de Modelos de *Forecasting* 22
Figura 2 – Boxplot de Volume 29
Figura 3 – Boxplot da Transformação de Volume com Box-Cox 30
Figura 4 – Frequência por Loja Atacadista (*Agency*) 30
Figura 5 – Frequência por Produto (SKU) 31
Figura 6 – Boxplot do preço 31
Figura 7 – Boxplot da temperatura 32
Figura 8 – Temperatura por Localização do Atacadista 33
Figura 9 – Desvio-Padrão da Temperatura por Localização do Atacadista 33
Figura 10 – População por Localização do Atacadista 34
Figura 11 – Renda Familiar Anual Média por Localização do Atacadista 35
Figura 12 – Boxplot do volume de vendas de refrigerantes 36
Figura 13 – Boxplot da produção de cerveja 36
Figura 14 – Evolução do Volume Global com o tempo 38
Figura 15 – Evolução do Preço Médio Global com o tempo 38
Figura 16 – Gráfico do Preço vs Volume 39
Figura 17 – Gráfico da Temperatura vs Volume 40
Figura 18 – Gráfico da Localização vs Volume 42
Figura 19 – Gráfico da População vs Volume 42
Figura 20 – Gráfico da Renda Familiar vs Volume 43
Figura 21 – Gráfico da Venda de Refrigerantes vs Volume 44
Figura 22 – Gráfico da Produção de Cerveja vs Volume 45
Figura 23 – Boxplots dos Eventos vs Volume 49
Figura 24 – Volume por SKU 49
Figura 25 – Volume por Loja Atacadista 50
Figura 26 – Matriz de Correlação de Pearson 56
Figura 27 – Gráfico do Método do Cotovelo para Agrupamento Loja-Produto . . . 57
Figura 28 – Comparativo do Volume por Grupo 59
Figura 29 – Comparativo do Preço por Grupo 60
Figura 30 – Comparativo da Temperatura da Região por Grupo 60
Figura 31 – Comparativo da População da Região por Grupo 61
Figura 32 – Comparativo da Renda da População da Região por Grupo 61
Figura 33 – Comparativo da Proximidade com o Litoral da Região por Grupo . . . 62
Figura 34 – Série Temporal Grupo A - Loja 23 - Produto 21 63
Figura 35 – Série Temporal Grupo B - Loja 26 - Produto 11 65
Figura 36 – Série Temporal Grupo C - Loja 60 - Produto 23 65

Figura 37 – Série Temporal Grupo D - Loja 32 - Produto 2 68
Figura 38 – Série Temporal Grupo E - Loja 2 - Produto 3 68

LISTA DE TABELAS

Tabela 1 – Arquivos da Base de Dados de Vendas de Cerveja 19
Tabela 2 – Atributos Preditivos da Base de Dados 20
Tabela 3 – Variável Resposta da Base de Dados 21
Tabela 4 – Exemplos de Dados Usados para Previsão de Vendas 21
Tabela 5 – Técnicas Aplicadas a Previsão de Vendas 23
Tabela 6 – Técnicas com os Melhores Desempenhos e Métricas Aplicadas 25
Tabela 7 – Localização das Lojas 34
Tabela 8 – Calendário de Eventos 37
Tabela 9 – Sobreposição das Lojas para as SKUs de Alto Volume 41
Tabela 10 – SKUs comercializadas pelas Lojas de Pequeno Volume 43
Tabela 11 – Características da Lojas de Pequeno Volume 44
Tabela 12 – Tabela de Modelos de Séries Temporais 46
Tabela 13 – Tabela de Combinação de Valores Extremos dos Atributos 46
Tabela 14 – Tabela dos Melhores Modelos segundo a Combinação de Atributos . . 47
Tabela 15 – Tabela das Melhores Métricas segundo a Combinação de Atributos . . 47
Tabela 16 – Combinação de Atacadistas (*Agency*) e SKU 48
Tabela 17 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 1 a 7 . . . 51
Tabela 18 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 8 a 14 . . . 52
Tabela 19 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 15 a 21 . . 53
Tabela 20 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 22 a 28 . . 54
Tabela 21 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 29 a 34 . . 55
Tabela 22 – Resumo do Perfil dos Grupos 59
Tabela 23 – Melhor Modelo de Previsão por Grupo 62
Tabela 24 – Grupo A - Previsões por Modelo 63
Tabela 25 – Grupo A - Desempenho por Modelo 64
Tabela 26 – Grupo B - Previsões por Modelo 64
Tabela 27 – Grupo B - Desempenho por Modelo 64
Tabela 28 – Grupo C - Previsões por Modelo 66
Tabela 29 – Grupo C - Desempenho por Modelo 66
Tabela 30 – Grupo D - Previsões por Modelo 67
Tabela 31 – Grupo D - Desempenho por Modelo 67
Tabela 32 – Grupo E - Previsões por Modelo 67
Tabela 33 – Grupo E - Desempenho por Modelo 69

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks*
AR	*Autoregressive*
ARI	*Autoregressive Integrated*
ARIMA	*Autoregressive Integrated Moving Average*
ARMA	*Autoregressive Moving Average*
BPN	*Back Propagation Neural Network*
CNN	*Convolutional Neural Network*
DNN	*Deep Neural Networks*
DT	*Decision Trees*
EAM	Erro Absoluto Médio
ELM	*Extreme Learning Machine*
EPAM	*Erro Percentual Absoluto Médio*
ETS	*Exponential Time Series Smoothing*
EQM	Erro Quadrático Médio
FA	*Forecast Accuracy*
FNN	*Feedforward Neural Network*
GBT	*Gradient Boosted Tree*
GHSOM	*Growing Hierarchical Self Organizing Map*
GLM	*Generalized Linear Model*
GMRAE	*Geometric Mean Relative Absolute Error*
HDVTEMP	*High*/Alto Desvio-Padrão da Temperatura
HINC	*High Income*/Alta Renda Familiar
hl	Hectolitros
HPOP	*High Population*/Alta População

HTEMP	*High Temperature*/Alta Temperatura
HVOL	*High Volume*/Alto Volume
KPI	*Key Performance Indicator*
LDVTEMP	*Low*/Baixo Desvio-Padrão da Temperatura
LINC	*Low Income*/Baixa Renda Familiar
LPOP	*Low Population*/Baixa População
LSTM	*Long Short-Term Memory*
LTEMP	*Low Temperature*/Baixa Temperatura
LVOL	*Low Volume*/Baixo Volume
MAD	*Mean Absolute Deviation*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MASE	*Mean Absolute Scaled Error*
MLR	*Multiple Linear Regression*
MSE	*Mean Squared Error*
MXN	Pesos Mexicanos
NN	*Neural Network*
NMI	*Normalized Mutual Information*
POCID	*Prediction of Change*
R2	Coeficiente de Determinação R2
REQM	Raiz do Erro Quadrático Médio
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
RMSPE	*Root Mean Square Percentage Error*
SARIMA	*Seasonal Autoregressive Integrated Moving Average*
SCM	*Supply Chain Management*

SKU	*Stock Keeping Unit*
SMA	*Simple Moving Average*
SMOreg	*Sequential Minimal Optimization for Regression*
SOM	*Self Organizing Map*
SVR	*Support Vector Regression*
TU	*Theil's U - uncertainty coefficient*
USD	Dólares Americanos

SUMÁRIO

1 INTRODUÇÃO 17
1.1 Importância do Tema 17
1.2 Objetivos do Projeto de Pesquisa 17
1.2.1 Objetivos Gerais 17
1.2.2 Objetivos Específicos 17
1.2.3 Base de Dados 18
1.3 Revisão Bibliográfica 18
1.3.1 Definição 18
1.3.2 Atributos para a Previsão de Vendas 18
1.3.3 Técnicas de Previsão de Vendas 21
1.3.4 Medidas de Desempenho 22
1.3.4.1 RMSE 24
1.3.4.2 MAPE 24
1.3.4.3 TU 26
1.3.4.4 POCID 26

2 METODOLOGIA 27
2.1 Fases da Metodologia 27
2.2 Coleta dos Dados 27
2.3 Pré-processamento 27
2.3.1 Tipos 27
2.3.2 Dados Faltantes 28
2.3.3 Dados Aberrantes 28
2.3.4 Transformação de Variáveis 28
2.3.5 Atributos Redundantes 28
2.3.6 Valores Errados, Duplicados e Organização dos Dados 29
2.4 Análise Exploratória de Dados - Univariada 29
2.4.1 Volume de Vendas de Cerveja 29
2.4.2 Lojas Atacadistas (*Agency*) 30
2.4.3 Produto (SKU) 30
2.4.4 Preço, Vendas Regulares & Promoção 31
2.4.5 Temperatura Máxima Média 32
2.4.6 Litoral 32
2.4.7 População Média 33
2.4.8 Renda Familiar Anual Média 34
2.4.9 Volume de Vendas de Refrigerantes 35

2.4.10 Produção de Cerveja . . . 35
2.4.11 Calendário de Eventos . . . 35
2.5 Análise Exploratória dos Dados - Multivariada . . . 37
2.5.1 Volume vs Mês-Ano . . . 37
2.5.2 Preço vs Mês-Ano . . . 37
2.5.3 Preço vs Volume . . . 38
2.5.4 Temperatura vs Volume . . . 39
2.5.5 Litoral vs Volume . . . 39
2.5.6 População vs Volume . . . 39
2.5.7 Renda Familiar vs Volume . . . 39
2.5.8 Vendas de Refrigerantes vs Volume . . . 39
2.5.9 Produção de Cerveja vs Volume . . . 40
2.5.10 Eventos vs Volume . . . 40
2.6 Análise Exploratória dos Dados - SKUs de Alto Volume . . . 40
2.7 Análise Exploratória dos Dados - Lojas de Baixo Volume . . . 42
2.8 Matriz de Correlação de Pearson . . . 43
2.9 Seleção dos Modelos para Ajuste Fino . . . 45
2.10 Agrupamento por Loja e Produto . . . 45
2.11 Ajuste fino dos Modelos Finais . . . 56

3 RESULTADOS . . . 59
3.1 Perfil dos Grupos . . . 59
3.2 Resumo dos Modelos por Grupo . . . 60
3.3 Resultados do Grupo A . . . 61
3.4 Resultados do Grupo B . . . 62
3.5 Resultados do Grupo C . . . 63
3.6 Resultados do Grupo D . . . 65
3.7 Resultados do Grupo E . . . 66

4 CONCLUSÃO . . . 71

REFERÊNCIAS . . . 73

1 INTRODUÇÃO

1.1 Importância do Tema

Uma previsão de vendas imprecisa prejudica a rentabilidade podendo trazer desde consequências leves como moderadas ou graves para um negócio (HUANG; FILDES; SOOPRAMANIEN, 2019). Se uma venda ocorre abaixo do previsto, há uma elevação do custo de estoque devido ao custo de capital, armazenagem e deterioração (COOPER et al., 1999). Quando falta um item especifico no estoque, há uma perda de venda e o lucro resultante. Se a situação se repete regularmente, pode levar a uma insatisfação do cliente que, no longo prazo, pode levá-lo a trocar pelo concorrente (CORSTEN; GRUEN, 2003).

Do ponto de vista matemático, modelos de previsão compreendem um tema abrangente que envolve diversas áreas da matemática, desde modelos de regressão linear simples até a utilização de processos estocásticos para a modelagem de séries temporais (GOOIJER; HYNDMAN, 2006).

Os avanços mais recentes em aquisição e capacidade de processamento de dados fazem com que diversos modelos, antes trabalhosos ou complexos demais, possam ser usados como técnicas de aprendizado de máquina aumentando o leque de possibilidades da Ciência de Dados para a obtenção de melhores resultados.

1.2 Objetivos do Projeto de Pesquisa

1.2.1 Objetivos Gerais

Propõe-se aplicar diversas técnicas de aprendizado de máquina para a criação de um modelo de previsão do volume de vendas de uma companhia produtora de cerveja.

1.2.2 Objetivos Específicos

- Estudar os modelos de regressão para séries temporais mais comumente empregados em *Sales Forecasting*;

- Comparar a performance dos modelos estudados utilizando códigos com a linguagem *Python* através de métricas como o RMSE(*Root Mean Squared Error*), MAPE(*Mean Absolute Percentage Error*), TU(*Theil's U*) - coeficiente de incerteza e POCID(*Prediction of Change*);

- Aplicar a metodologia em um conjunto de dados real.

1.2.3 Base de Dados

Serão usados os dados públicos da ABInBev, uma empresa multinacional de bebidas e cervejas formada em 2004 pela fusão da belga Interbrew e da brasileira Ambev Os dados estão disponíveis em: https://www.kaggle.com/utathya/future-volume-prediction, e estão contidos em 7 arquivos com detalhes por produto (*SKU*) e por loja atacadista (*Agency*) conforme mostrado na Tabela 1 com suas descrições, unidade de medida, número de amostras e número de atributos, sendo que a variável que se deseja prever é o volume de vendas de cerveja em hectolitros (hl).

Os detalhes dos atributos preditivos e suas principais características são mostrados na Tabela 2.

Os detalhes da variável resposta Volume correspondem aos hectolitros de vendas de cerveja e suas principais características são mostradas na Tabela 3.

1.3 Revisão Bibliográfica

1.3.1 Definição

Sales Forecasting ou Previsão de Vendas pode ser definida como uma projeção para o futuro da demanda esperada, dado um conjunto declarado de condições ambientais (MENTZER; MOON, 2004a).

Esse é um dos processos mais importantes dentro de uma organização pois seu resultado pode impactar as ações de todas as áreas, tanto táticas quanto estratégicas. Além disso, é um dos mais complexos, pois envolve uma série de componentes, como ambiente, mercado, ações da empresa, de competidores, de fornecedores, de distribuidores, do governo, custos, lucratividade e *market-share* (ARMSTRONG, 2001).

1.3.2 Atributos para a Previsão de Vendas

Os componentes que influenciam a previsão de vendas também podem ser chamados de atributos ou variáveis e podem mudar conforme o problema a ser analisado, seja para as vendas do varejo, de alimentos, computadores, bebidas, transportes, produtos da indústria têxtil, farmacêutica ou química entre outras.

Podem ser classificados em qualitativas e quantitativas (MENTZER; MOON, 2004b). As variáveis qualitativas têm seu impacto julgado por especialistas com base em suas experiências e representam uma opinião ou julgamento, sendo muito valiosos em muitas situações. Já as variáveis quantitativas representam um conjunto de dados históricos sobre as vendas realizadas e podem conter informação sobre o produto, família de produtos, loja, preço, localização geográfica, localização de competidores entre outros.

Singh et al. (2018) por exemplo usam o atributo número de passageiros para estudar

Tabela 1 – Arquivos da Base de Dados de Vendas de Cerveja

arquivo.csv	**Descrição**	**Unidade**	**#Amostras**	**#Atributos**
price_sales_promotion	Holds the price, sales & promotion in dollar value per hectoliter at Agency-SKU-month level	USD/ hectoliter	21000	6
historical_volume	Holds sales data at Agency-SKU-month level from Jan 2013 to Dec 2017	hectoliters	21000	5
weather	Holds average maximum temperature at Agency-month level	Degree Celsius	3600	3
industry_soda_sales	Holds industry level soda sales	hectoliters	60	2
event_calendar	Holds event details (sports, carnivals, etc.)	-	61	13
industry_volume	Holds industry actual beer volume	hectoliters	60	2
demographics	Holds demographic details (Yearly income in USD)	-	60	3

Fonte: Elaborada pelo autor.

Nota: Extraído da base de dados.

Tabela 2 – Atributos Preditivos da Base de Dados

Atributo	**Descrição**	**#Val**	**Faixa de Valores**	**T**
Agency	Wholesalers id	60	Agency_01-60	O
Avg_Population_2017	Avg pop/Agcy, 2017	60	12271-3137874	I
Avg_Yearly_Household _Income_2017	Avg household income, USD/year, 2017	60	90240-247220	I
YearMonth	Calendar	60	1/2013-12/2017	I
Easter Day	Event or Holiday	5	3/13,4/14-15-16-17	I
Good Friday	Event or Holiday	5	3/13,4/14-15-16-17	I
New Year	Event or Holiday	5	Every January,1	I
Christmas	Event or Holiday	5	Every Decemb,25	I
Labor Day	Event or Holiday	5	Every May,1	I
Independency Day	Event or Holiday	5	Every September	I
Revolution Day Memorial	Event or Holiday	5	Every November	I
Regional Games	Juegos Centroamer. y del Caribe	1	Every 4 years, 11/2014	I
FIFA U-17 World Cup	FIFA U-17 World Cup	0	No ocurrence	I
Football Gold Cup	CONCACAF Cup	0	No ocurrence	I
Beer Capital	Event or Holiday (yearly)	5	10/13-14, 11/15-16-17	I
Music Fest	Event or Holiday (yearly)	5	3/13-14-15, 4/16, 3/2017	I
SKU	Stock Keeping Unit	25	SKU_01 to 34	O
Soda_Volume	Soda Sales (hectoliters)	60	696401477-1049868815	I
Industry_Volume	Beer Production (hectoliters)	60	413051813-670015726	I
Price (Beer)	Price: Sales+Prom (USD/hl/agcy)	16727	0.000-19166.625	F
Sales (Beer)	Regular Price (USD/hl/agcy)	17962	(3121.690)-4925.404	F
Promotions (Beer)	Promo Price (USD/hl/agcy)	17511	0.000-19166.625	F
Avg_Max_Temp	Avg Max Temp (oC)	1418	16.731-45.290	F

Fonte: Elaborada pelo autor.

Nota: Extraído da base de dados.

Tabela 3 – Variável Resposta da Base de Dados

Atributo	Descrição	#Val	Faixa de Valores	T
Volume	Beer Sales hectoliters at Agency-SKU-month level	14098	0.00-22526.61	F

Fonte: Elaborada pelo autor.

Nota: Extraído da base de dados.

companhias aéreas internacionais e diferentes SKUs para os vinhos australianos enquanto que Sagaert et al. (2018) incorporaram indicadores macroeconômicos para reduzir o erro dos modelos.

A tabela Tabela 4 mostra os dados usados que foram encontrados na pesquisa bibliográfica efetuada para esse trabalho.

Tabela 4 – Exemplos de Dados Usados para Previsão de Vendas

Autor(es)	Dados Usados
Singh et al. (2018)	Airlines (número de passageiros) Wine (vinhos australianos - fortificado, branco seco, branco doce, vermelho, rosa e cintilante)
Chen e Lu (2017)	Computadores (PC-Personal Computer), NB-Notebook e LCD-Liquid Crystal Display)
Pavlyshenko (2019)	Rossman (vendas das lojas)
Loureiro, Miguéis e Silva (2018)	Moda (varejo de roupas de moda)
Cheriyan et al. (2018)	E-Moda (varejo de roupas de moda pela internet)
Jiang et al. (2020)	Vodka (duas marcas de um distribuidor de bebidas)
Kuo, Tseng e Chen (2016)	Laptop (distribuidor de Taiwan)
Sagaert et al. (2018)	Pneus (matéria-prima para pneus de veículos de passeio e caminhões)
Krishna et al. (2018)	Varejo (loja de varejo)
Merkuryeva, Valberga e Smirnov (2019)	Farmacêutico (um produto)

Fonte: Elaborada pelo autor.

Nota: Baseada na pesquisa bibliográfica sobre Previsão de Vendas.

1.3.3 Técnicas de Previsão de Vendas

Herbig, Milewicz e Golden (1993) explicam que as técnicas de previsão variam de simples a complexas, podendo existir combinações para a construção de novos modelos

e que todas são concebidas para produzir estimativas precisas e imparciais de previsões futuras na presença de incerteza.

De acordo com Gallagher, Madden e D'Arcy (2015), o aprendizado de máquina e as estatísticas são amplamente usados para a análise de vendas e para determinar os fatores significativos na previsão de vendas. Alguns dos métodos usados nas indústrias incluem regressão linear, redes neurais, redes Bayesianas, regressão multivariada e séries temporais.

Muitas técnicas diferentes são usadas e uma visão geral do resultado da pesquisa bibliográfica efetuada é mostrada na tabela Tabela 5.

A prática mostra que o uso de abordagens de regressão pode frequentemente nos dar melhores resultados em comparação com métodos de séries temporais (PAVLYSHENKO, 2019).

1.3.4 Medidas de Desempenho

As métricas de avaliação do desempenho do modelo devem ser feitas sobre o subconjunto de teste obtido do conjunto total de dados, conforme Figura 1.

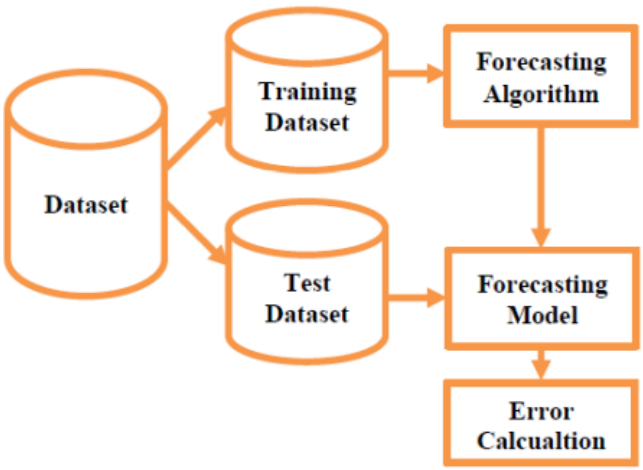

Figura 1 – Avaliação de Modelos de *Forecasting*

Fonte: (SINGH et al., 2018)

Um "erro" de previsão é a diferença entre um valor observado e sua previsão. Aqui, "erro" não significa um erro em si, mas sim a parte imprevisível de uma observação. Pode ser escrito como:

$$Erro = y_t - \hat{y}_t$$

onde os dados de teste são fornecidos por y_t, os valores previstos são $\hat{y}_t$ e T é o tamanho dos dados de teste com t = { 1, ..., T }.

Tabela 5 – Técnicas Aplicadas a Previsão de Vendas

Sigla	**Descrição**
AdaB	AdaBoost
ANN	Artificial Neural Networks
AR	Autoregressive
ARI	Autoregressive Integrated
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BPN	Back Propagation Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Networks
DT	Decision Trees
ELM	Extreme Learning Machine
ET	Extra Tree
ETS	Exponential Time Series Smoothing
FNN	Feedforward Neural Network
GBT	Gradient Boosted Tree
GHSOM	Growing Hierarchical Self Organizing Map
GLM	Generalized Linear Model
GM	Gaussian Model
GradB	GradientBoost
HW	Holt-Winters
KMeans	K-Means
LR	Linear Regression
LSTM	Long Short-Term Memory
MLP	Multilayer perceptron
MLR	Multiple Linear Regression
NN	Neural Network
RF	Random Forest
RLasso	Lasso Regression
RPoly	Polynomial Regression
RRidge	Ridge Regression
RStep	Step Regression
RSymb	Symbolic Regression
SMA	Simple Moving Average
SMOreg	Sequential Minimal Optimization for Regression
SOM	Self Organizing Map
STK	Stacking
SVR	Support Vector Regression

Fonte: Elaborada pelo autor.

Nota: Baseada na pesquisa bibliográfica sobre Previsão de Vendas.

Observe-se que os erros de previsão são diferentes dos resíduos de duas maneiras. Primeiro, os resíduos são calculados no conjunto de treinamento, enquanto os erros de previsão são calculados no conjunto de teste. Em segundo lugar, os resíduos são baseados em previsões de uma etapa, enquanto os erros de previsão podem envolver várias etapas (HYNDMAN; ATHANASOPOULOS, 2018).

A Tabela 6 mostra as diferentes métricas utilizadas e os melhores modelos para cada base de dados encontrados na pesquisa bibliográfica efetuada.

Adotaremos como principais métricas para este estudo, o RMSE, o MAPE, o TU e o POCID. O RMSE porque possui a mesma unidade de medida da variável sendo predita o que facilita a sua interpretação, o MAPE porque a expressão dos resultados em porcentagem também facilita muito a compreensão além do que o seu complementar, FA, é muito comumente utilizado pelos profissionais de *Supply Chain*, e finalmente o TU e POCID que são métricas específicas para o estudo de séries temporais.

1.3.4.1 RMSE

Root Mean Square Error (RMSE) ou Raiz do Erro Quadrático Médio (REQM). É uma medida do erro do modelo onde n é o número de dados, y_j é o vetor de valores observados e $\hat{y}_j$ é o vetor de valores preditos, para cada uma das n observações j = { 1, ..., n }:

$$RMSE = \frac{1}{n} * \sqrt{\sum_{j=1}^{j=n} (y_j - \hat{y}_j)^2}$$

1.3.4.2 MAPE

Mean Absolute Percentage Error (MAPE) ou Erro Percentual Absoluto Médio (EPAM). É uma medida do erro do modelo onde y_j é o valor real e $\hat{y}_j$ é o valor predito. É um indicador simples, porém seu uso tem algumas limitações:

- Quando existem valores de y_j=0 não pode ser usado porque haveria uma divisão por zero.
- Para diferenças grandes entre y_j e $\hat{y}_j$, o indicador poderá ser maior que 100% e sem limite superior.
- Quando $y_j < \hat{y}_j$, há uma penalização maior para os erros negativos do que para os positivos tornando-o viesado e favorecendo os resultados com os $\hat{y}_j$ mais baixos. Uma alternativa para evitar isso é usar o *log(previsto/real)*

$$MAPE = 100/n * \sum_{j=1}^{j=n} |(y_j - \hat{y}_j) / y_j|$$

Tabela 6 – Técnicas com os Melhores Desempenhos e Métricas Aplicadas

Base de Dados	**Melhor Resultado**	**Outras Técnicas Utilizadas**	**Métricas**
Airlines Wine	SMOreg (Airlines) LR (Wine)	GM MLP	MAE RMSE
Computadores	GHSOM+ELM	SVR ELM GHSOM+SVR KMeans+SVR SOM+SVR KMeans+ELM SOM+ELM	MAPE RMSPE
Rossman	STK	RF ET ARIMA RLasso NN	MAE/ (Vendas)
Moda	DNN	DT RF SVR ANN LR	R2 RMSE MAPE MAE MSE
E-Moda	GBT	GLM DT	RMSE MSE MAE
Vodka	ARMA ARIMA(d=1)	Naive1,Naive2 MLR biLSTM CNN+LSTM LSTM	MAE
Laptop	BPN+FNN	FNN	MSE
Pneus	RLasso	Naive1,Naive2 HW ETS ARIMA ARI LR RStep RLasso+AR	MAPE MASE GMRAE
Varejo	GradB	LR RPoly RLasso RRidge AdaB	RMSE R2
Farmacêutico	MLR RSymb	SMA	MAD MAE

Fonte: Elaborada pelo autor.

Nota: Baseada na pesquisa bibliográfica sobre Previsão de Vendas.

O seu complementar, chamado de FA (*forecast accuracy*) é um KPI do processo de *Supply Chain Management* (SCM).

$$FA = 1 - MAPE$$

1.3.4.3 TU

Theil's U (TU) também chamado de coeficiente de incerteza, proficiência ou coeficiente de entropia. Foi primeiramente introduzido por Henri Theil e é baseado no conceito da entropia da informação. Mede o quão bom é o modelo de previsão em relação a um modelo de referência ou *baseline*, em geral um modelo *naive*.

Interpretação:

- TU <1: a técnica de previsão é melhor do que a *naive*.
- TU=1: a técnica de previsão é tão boa quanto a *naive*.
- TU>1: a técnica de previsão é pior do que a *naive*.

$TU = \sum_{t=1}^{h} (z_t - \hat{z}_t)^2 / \sum_{t=1}^{h} (z_t - z_{t-1})^2$ Onde: z_t e z_{t-1} são os valores reais de um ponto no tempo t e t-1, e $\hat{z}_t$ é o valor previsto.

1.3.4.4 POCID

Prediction of Change (POCID) é a porcentagem de predição correta de aumento ou diminuição do valor da série. Portanto, quanto mais próximo de 100%, melhor é a previsão.

POCID = $\frac{\sum_{t=1}^{h} D_t}{h} * 100$

Onde: $D = 1 se (\hat{z}_t - \hat{z}_{t-1}) * (z_t - z_{t-1}) > 0 D = 0 se (\hat{z}_t - \hat{z}_{t-1}) * (z_t - z_{t-1}) < 0$

2 METODOLOGIA

2.1 Fases da Metodologia

Neste trabalho, seguiremos algumas fases principais para o desenvolvimento do projeto, a saber:

- Coleta dos Dados;
- Pré-processamento;
- Análise Exploratória de Dados - Univariada;
- Análise Exploratória de Dados - Multivariada;
- Seleção dos Modelos;
- Agrupamento por Loja e Produto;
- Ajuste fino do Modelo Final.

2.2 Coleta dos Dados

Consiste em agrupar as diferentes bases de dados em uma base única.

As 7 diferentes bases de dados foram agrupadas mantendo a granularidade mensal por Atacadista (Agency) e por SKU.

Os demais atributos preditivos, temperatura, população, renda familiar, vendas de refrigerantes, produção de cerveja e as datas festivas e feriados foram organizados conforme a estratégia acima, bem como o atributo objetivo volume de vendas de cerveja.

2.3 Pré-processamento

Esta é a etapa de preparação dos dados para posterior exploração e modelagem.

2.3.1 Tipos

O atributo preditivo *YearMonth* apresentava-se originalmente como objeto e foi alterado para o tipo *datetime*.

Os atributos *Agency* e *SKU* apresentavam-se como objeto e foram alterados para o tipo inteiro.

2.3.2 Dados Faltantes

Ocorreram casos de dados faltantes (*missing*) na base final única construída nos atributos de População e Renda Familiar pois os dados originais continham apenas um valor único para cada região e apenas para o ano de 2017.

Para o atributo população, foi obtida a população dos anos faltantes e a mensualização foi obtida por interpolação. Detalhes são mostrados na subseção 2.4.7.

Para o atributo renda familiar em moeda local, os dados encontrados na pesquisa efetuada foram da renda familiar per capita em dólares americanos e apenas para os anos pares. Foi realizada a conversão de *USD* para a moeda local, calculada a proporção entre a renda familiar per capita e a renda familiar, e feita a interpolação para preencher os dados faltantes. Detalhes são mostrados na subseção 2.4.8.

2.3.3 Dados Aberrantes

Dados aberrantes (*outliers*) foram encontrados em alguns atributos: Volume (subseção 2.4.1), Preço (subseção 2.4.4) e Temperatura (subseção 2.4.5) usando a ferramenta boxplot.

Porém, em nenhum dos casos existe informação de erro ou outra qualquer que permita sua remoção ou correção, e, portanto os dados serão mantidos tais quais na base de dados.

2.3.4 Transformação de Variáveis

Foram estudadas algumas transformações (*feature engineering*) para o atributo objetivo Volume (subseção 2.4.1) no sentido de corrigir a assimetria: logarítmica, raiz quadrada, 1/(x+c), boxcox, e *powertransformer* com box-cox e com yeo-johnson, porém nenhuma delas conseguiu corrigir a alta frequência de valores nulos e próximos de zero.

Foi criado um novo atributo 'Litoral' (*Coast*, subseção 2.4.6) para identificar se a loja atacadista se localiza numa região de litoral ou interior. A criação dessa variável foi obtida através do desvio-padrão da variável temperatura, sabendo-se que regiões do litoral possuem um menor desvio-padrão.

2.3.5 Atributos Redundantes

O preço (*price*, subseção 2.4.4) representa a soma do preço regular (*sales*) com as vendas promocionais (*promotion*).

Foi utilizado apenas o atributo preço para efeito dos estudos de modelagem, pois é esse total que influencia a variação do Volume.

2.3.6 Valores Errados, Duplicados e Organização dos Dados

Não foram encontrados dados preenchidos incorretamente ou duplicados nem houve necessidade de organizar corretamente os dados por linhas e colunas.

2.4 Análise Exploratória de Dados - Univariada

2.4.1 Volume de Vendas de Cerveja

O volume de vendas de cerveja em hectolitros apresenta 12,1% de dados com volume igual a zero no nível *Agency-SKU*, excluindo as SKUs que nunca tiveram venda em determinada Agency.

Os dados apresentam uma forte assimetria positiva e uma grande quantidade de pontos classificados estatisticamente como *outliers* conforme a Figura 2.

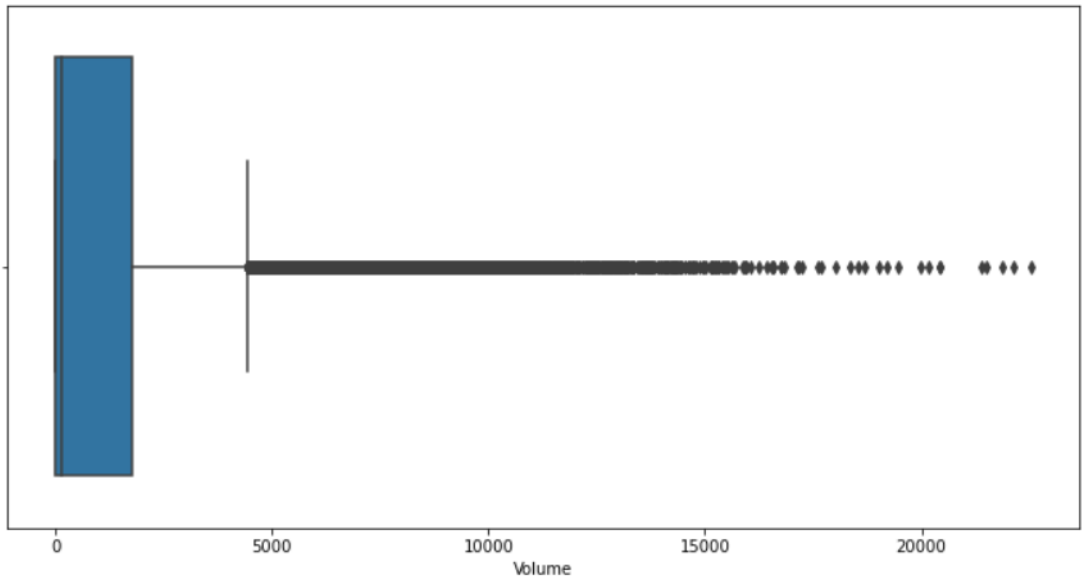

Figura 2 – Boxplot de Volume

Fonte: Elaborada pelo autor

Não temos informações que permitam tratar ou remover os *outliers*.

Quanto à assimetria, algumas transformações foram tentadas, como: logarítmica, raiz quadrada, 1/(x+c), boxcox, e *powertransformer* com box-cox e com yeo-johnson.

A transformação que melhor apresentou resultados foi a de boxcox, conforme a Figura 3, embora a grande quantidade de volumes iguais a zero ainda causem uma assimetria forte.

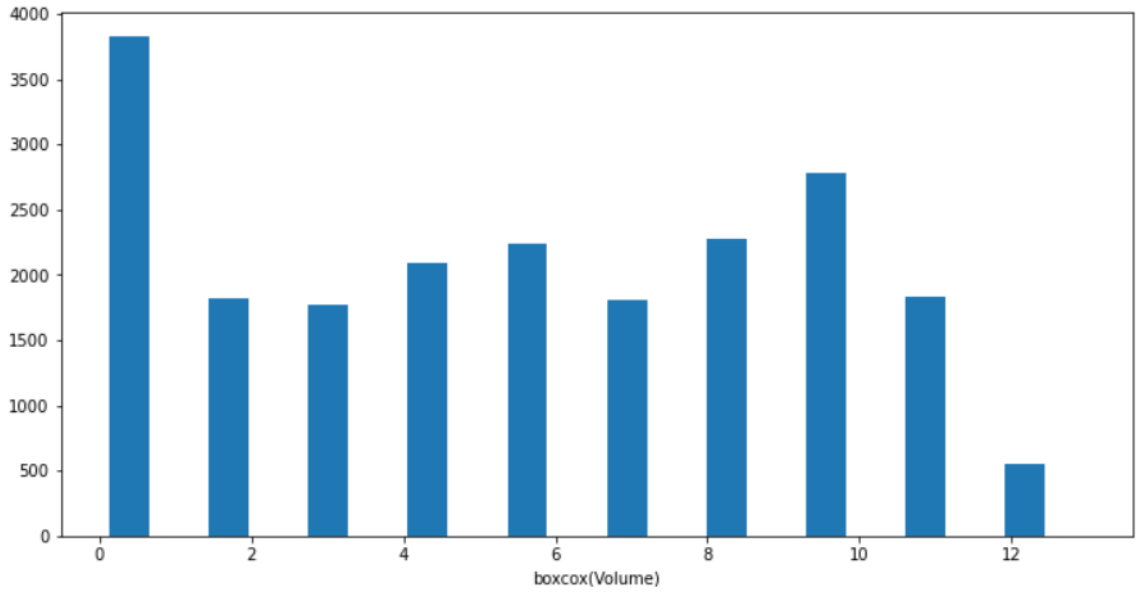

Figura 3 – Boxplot da Transformação de Volume com Box-Cox

Fonte: Elaborada pelo autor

2.4.2 Lojas Atacadistas (*Agency*)

A frequência de dados dos atacadistas (*Agency*) não apresenta uma grande diferença entre elas, indo de um patamar mínimo próximo a 100 a um máximo próximo a 500, e pode ser verificado na Figura 4.

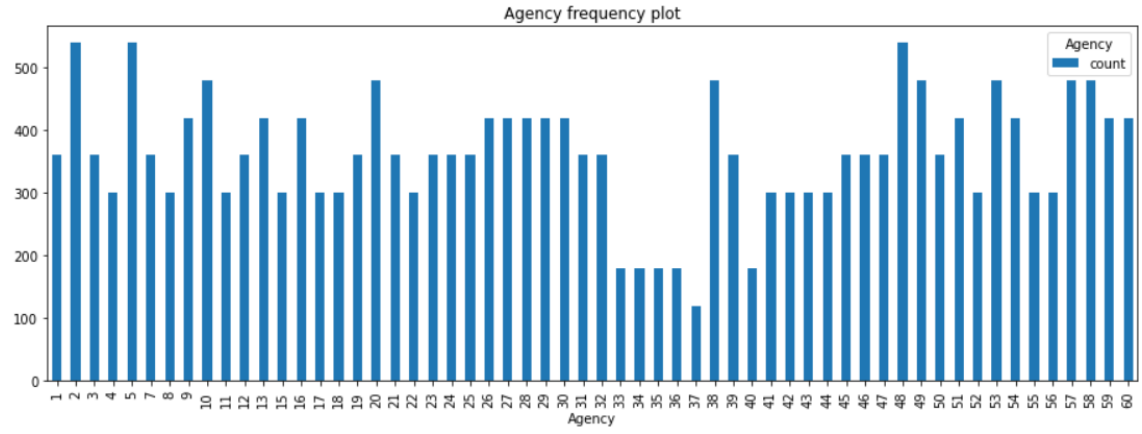

Figura 4 – Frequência por Loja Atacadista (*Agency*)

Fonte: Elaborada pelo autor

2.4.3 Produto (SKU)

As SKUs com maior frequência são as de número 1 até 5, representando em conjunto 77,4% do total. Todas as demais têm uma baixa frequência. Vide Figura 5.

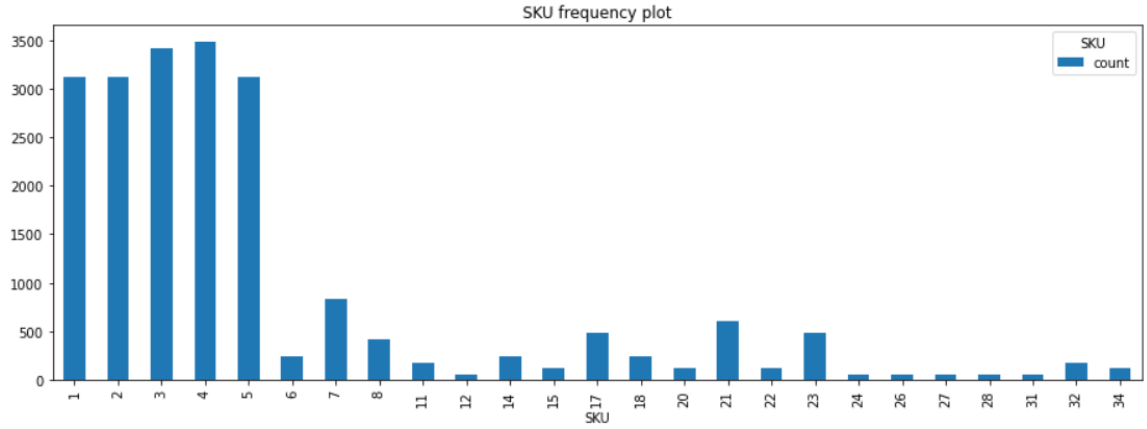

Figura 5 – Frequência por Produto (SKU)

Fonte: Elaborada pelo autor

2.4.4 Preço, Vendas Regulares & Promoção

O preço (*price*) representa a soma, em usd por hectolitros, do preço regular (*sales*) com as vendas promocionais (*promotion*).

O boxplot de *price* pode ser visto na Figura 6.

Existem 2 dados discrepantes com valor extremamente elevado (> 6000) e eles são devido a duas vendas promocionais a volumes extremamente baixos e próximos de zero.

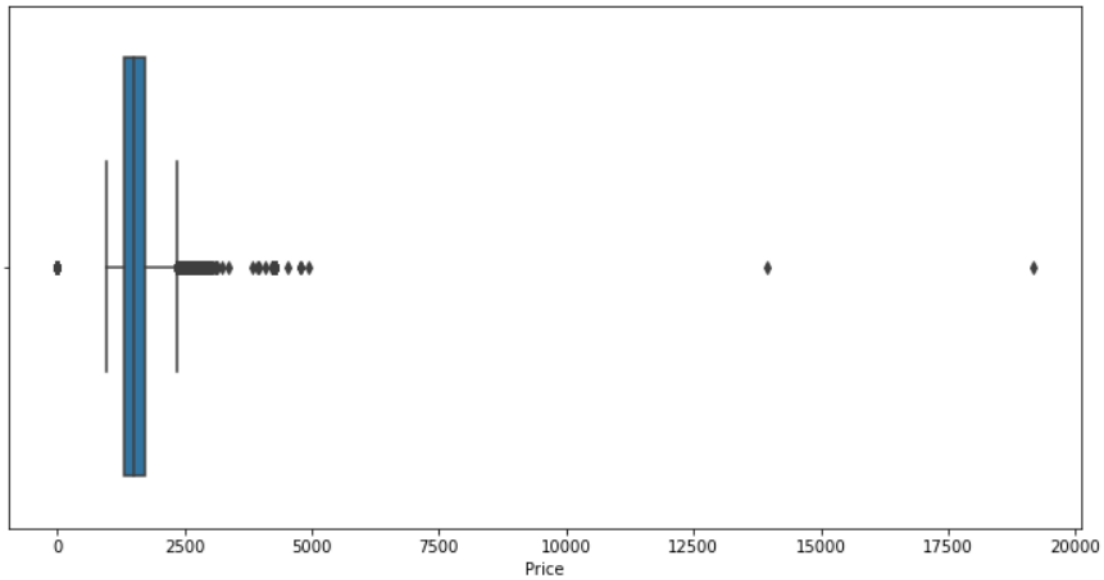

Figura 6 – Boxplot do preço

Fonte: Elaborada pelo autor

2.4.5 Temperatura Máxima Média

A temperatura máxima média (*Average Maximum Temperature*) varia de um mínimo de 17 a um máximo de 45 ^oC e não apresenta assimetria. O boxplot é mostrado na Figura 7.

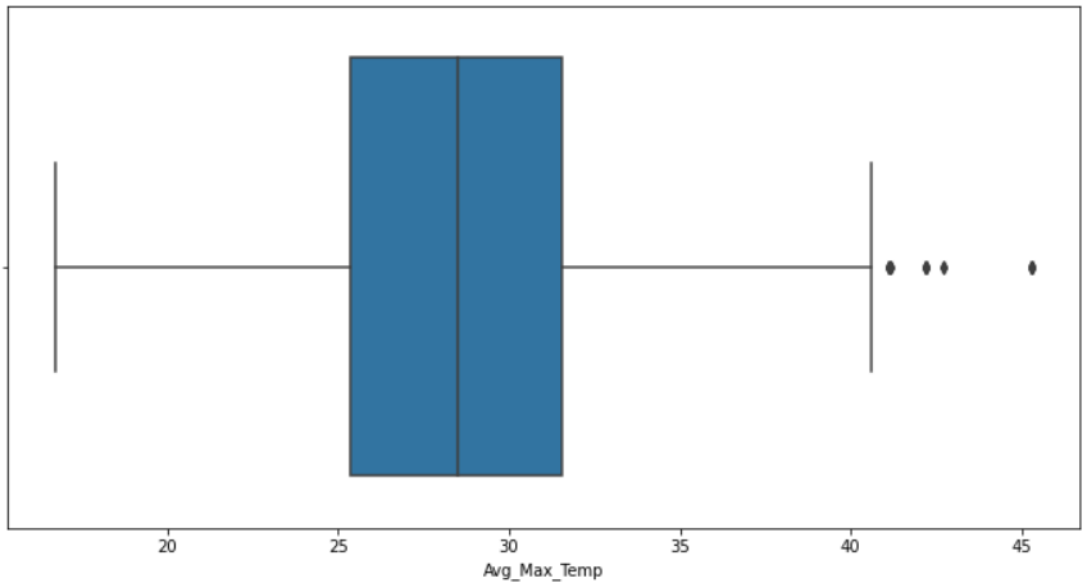

Figura 7 – Boxplot da temperatura

Fonte: Elaborada pelo autor

As Agency localizadas em regiões com as maiores temperaturas máximas médias são, na ordem decrescente: 37, 28, 45, 9 e 47. Já as menores temperaturas foram encontradas nas regiões: 60, 59, 13, 16, 58, 57, 38, 20, 39, 17, 12, 15, todas com os mesmos valores. A Figura 8 mostra o perfil de temperatura por Agency.

2.4.6 Litoral

Um outro aspecto considerado é se a localização da loja fica no litoral ou no continente. Apesar de não conhecermos essa localização geográfica, podemos inferir a partir do desvio-padrão da temperatura, pois cidades localizadas no litoral têm um menor desvio-padrão e as cidades localizadas no continente o oposto. As regiões com o menor desvio-padrão de temperatura (maior chance de serem litorâneas) na ordem crescente são: 9, 47, 45, 31 e 52. E os maiores desvio-padrão (maior chance da região ser continental) são, por ordem decrescente: 26, 4, 55, 56 e 3. A Figura 9 mostra o perfil do desvio-padrão da temperatura por Agency.

A classificação para cidades no litoral respeitou o limite superior de 2,00 oC como desvio-padrão para a temperatura, e o resultado é mostrado na Tabela 7.

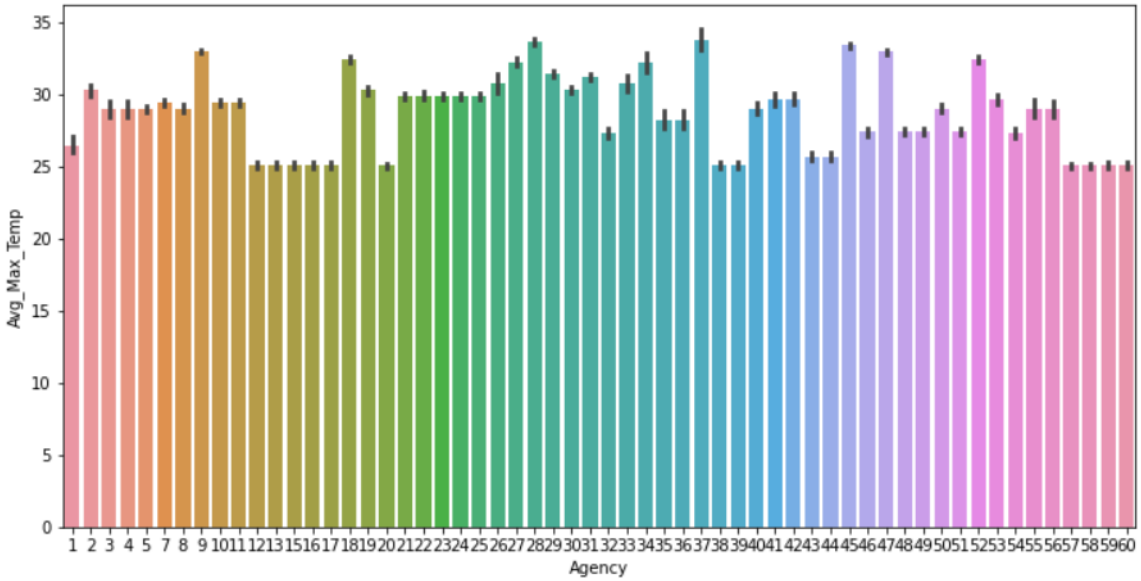

Figura 8 – Temperatura por Localização do Atacadista

Fonte: Elaborada pelo autor

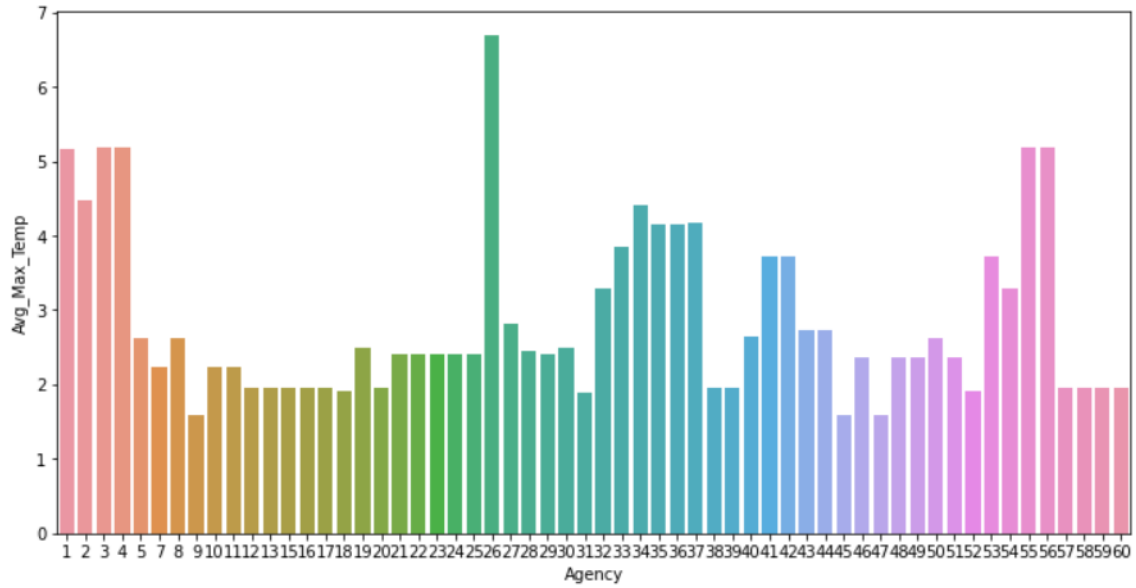

Figura 9 – Desvio-Padrão da Temperatura por Localização do Atacadista

Fonte: Elaborada pelo autor

2.4.7 População Média

A população média (*Average Population*) fornecida refere-se apenas à média anual do ano de 2017 e a sua distribuição por Agency é dada pela Figura 10.

As Agency localizadas em regiões mais populosas são, na ordem decrescente: 2, 5, 57, 60 e 6, e as menos populosas em ordem crescente são: 34, 36, 35, 21 e 29.

Tabela 7 – Localização das Lojas

Localização	Qtidade	Lojas
Litoral	18	9,12,13,15,16,17,18,20,31,38,39,45,47,52,57
		58,59,60
Interior	40	1,2,3,4,5,7,8,10,11,19,21,22,23,24,25,26,27,
		28,29,30,32,33,34,35,36,37,40,41,42,43,44,
		46,48,49,50,51,53,54,55,56

Fonte: Elaborada pelo autor.

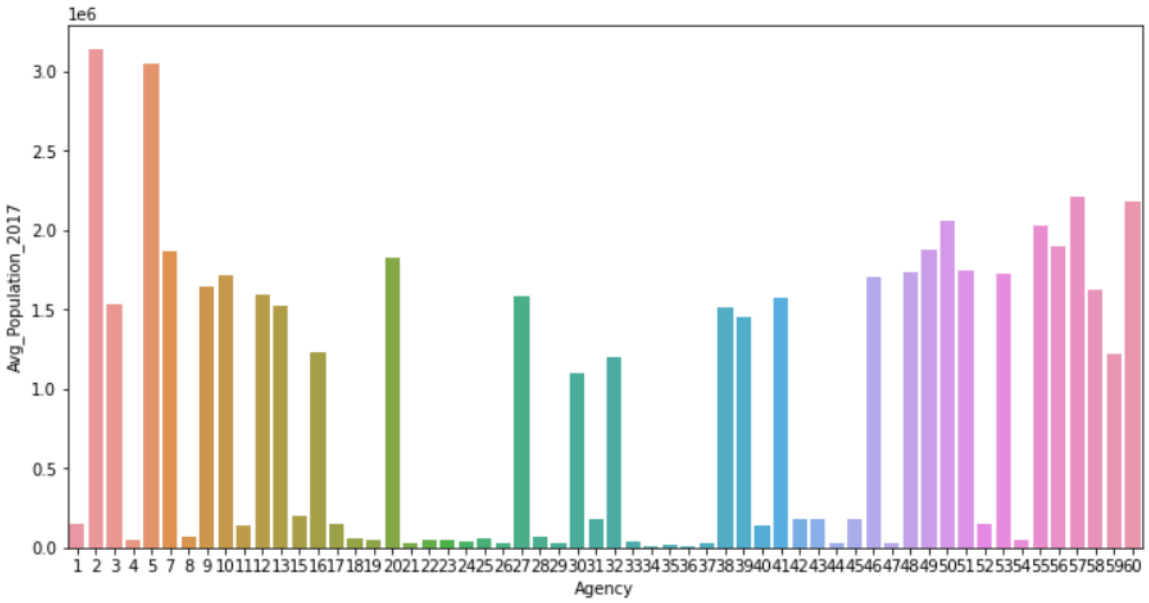

Figura 10 – População por Localização do Atacadista

Fonte: Elaborada pelo autor

Considerando que o calendário de feriados e eventos coincide com o México, podemos estimar que a soma da população de 2017 de todas as regiões corresponde a 56,2 milhões de pessoas ou 45 % da população total do México.

Foi feita a interpolação dos dados com a população anual do México de modo a se obter a cada mês no período de 2013 a 2017 os valores por atacadista.

2.4.8 Renda Familiar Anual Média

A renda familiar anual média (*Average Yearly Household Income*) fornecida pelo conjunto de dados refere-se apenas à média anual do ano de 2017 e a sua distribuição por Agency é dada pela Figura 11.

As Agency localizadas em regiões de maior renda familiar são, na ordem decrescente: 30, 16, 2, 51 e 55, e as de menor renda em ordem crescente são: 23, 26, 33, 24 e 35.

A renda familiar média de 2017 das regiões estudadas é de 148.119 pesos mexicanos

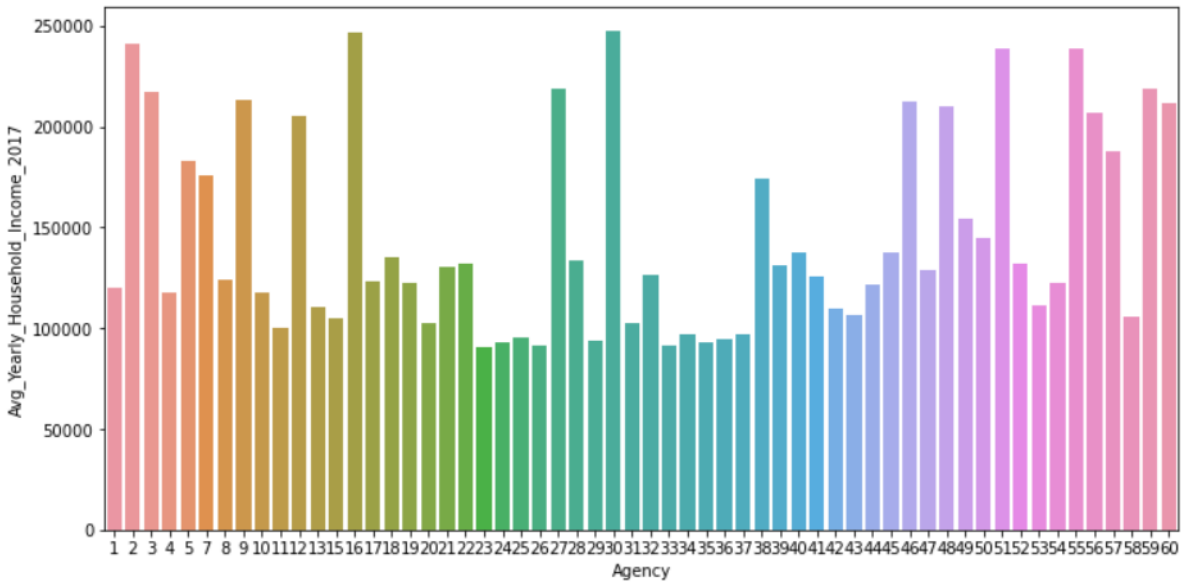

Figura 11 – Renda Familiar Anual Média por Localização do Atacadista

Fonte: Elaborada pelo autor

por ano, o que significa que é 2,76 maior que renda familiar média do país inteiro.

Para a interpolação dos dados e obtenção dos valores mensais para o período de 2013 a 2017 foram utilizados os dados de renda familiar média per capita em dólares americanos disponíveis em https://www.ceicdata.com/en/indicator/mexico/annual-household-income-per-capita (consultado em 20/11/2021).

2.4.9 Volume de Vendas de Refrigerantes

O volume de vendas de refrigerantes em hectolitros (*Soda Volume*) tem o seu boxplot apresentado na Figura 12 e não apresenta *outliers*.

O volume de vendas de cerveja dos atacadistas corresponde a apenas 0,0002 % do volume de vendas de refrigerantes.

2.4.10 Produção de Cerveja

O boxplot da produção de cerveja em hectolitros (*Industry Volume*) é apresentado na Figura 13 e não apresenta *outliers*.

O volume de vendas de cerveja dos atacadistas corresponde a 0,0003 % do volume de produção de cerveja.

2.4.11 Calendário de Eventos

O calendário de feriados e eventos permitiu concluir que o país objeto do estudo é o México.

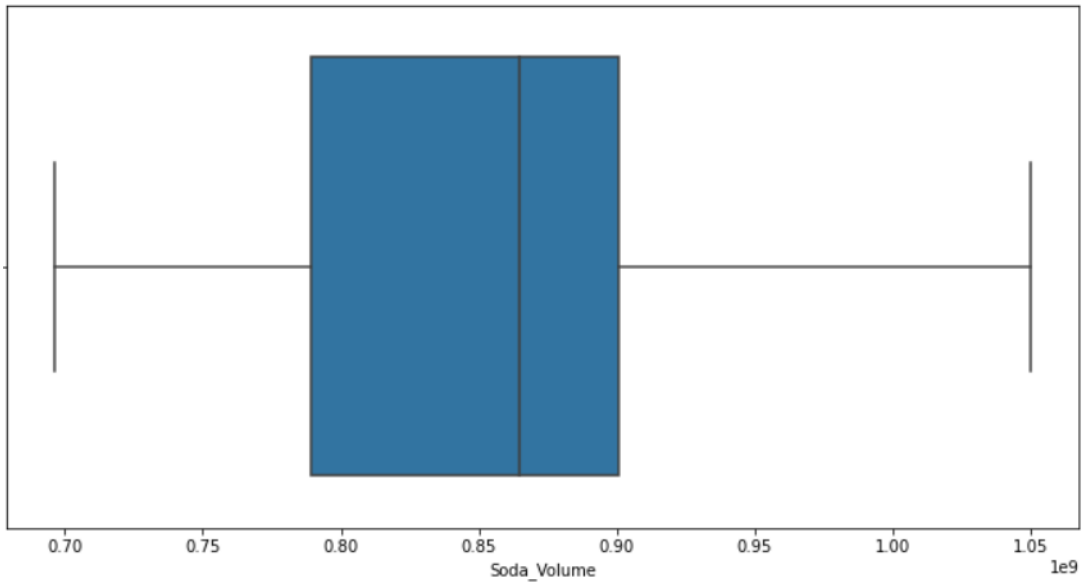

Figura 12 – Boxplot do volume de vendas de refrigerantes

Fonte: Elaborada pelo autor

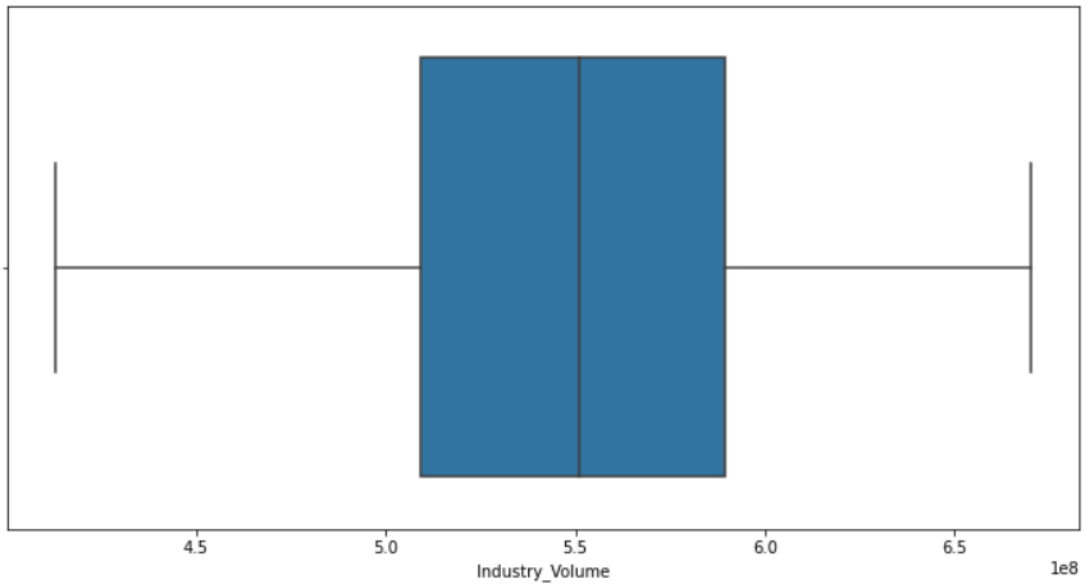

Figura 13 – Boxplot da produção de cerveja

Fonte: Elaborada pelo autor

Não foi possível, no entanto, determinar se a distribuição geográfica corresponde a todo o país ou a apenas uma região específica.

Os detalhes com as informações disponíveis são mostrados na Tabela 8

Tabela 8 – Calendário de Eventos

Evento	Descrição	#Val	Faixa de Valores	T
Easter Day	Event or Holiday	5	3/13,4/14-15-16-17	I
Good Friday	Event or Holiday	5	3/13,4/14-15-16-17	I
New Year	Event or Holiday	5	Every January,1	I
Christmas	Event or Holiday	5	Every Decemb,25	I
Labor Day	Event or Holiday	5	Every May,1	I
Independency Day	Event or Holiday	5	Every September	I
Revolution Day Memorial	Event or Holiday	5	Every November	I
Regional Games	Juegos Centroamer. y del Caribe	1	Every 4 years, 11/2014	I
FIFA U-17 World Cup	FIFA U-17 World Cup	0	No ocurrence	I
Football Gold Cup	CONCACAF Cup	0	No ocurrence	I
Beer Capital	Event or Holiday (yearly)	5	10/13-14, 11/15-16-17	I
Music Fest	Event or Holiday (yearly)	5	3/13-14-15, 4/16, 3/2017	I

Fonte: Elaborada pelo autor.

Nota: Extraído da base de dados.

2.5 Análise Exploratória dos Dados - Multivariada

Consiste em explorar os diferentes atributos preditivos e o atributo objetivo buscando encontrar associação entre os mesmos.

Não é uma fase definitiva e todas as pistas terão que ser validadas posteriormente, já que o objetivo maior nessa fase é uma melhor compreensão dos dados que estão sendo estudados.

2.5.1 Volume vs Mês-Ano

A série temporal do Volume Global (soma de todos os volumes) parece não ter tendência, mas parece ter uma sazonalidade bem definida, com vales todo início de ano, porém os picos variam de ano para ano. Vide Figura 14.

2.5.2 Preço vs Mês-Ano

A evolução global do preço demonstra uma tendência positiva, exceto nos últimos 3 meses quando a tendência é negativa, conforme a Figura 15.

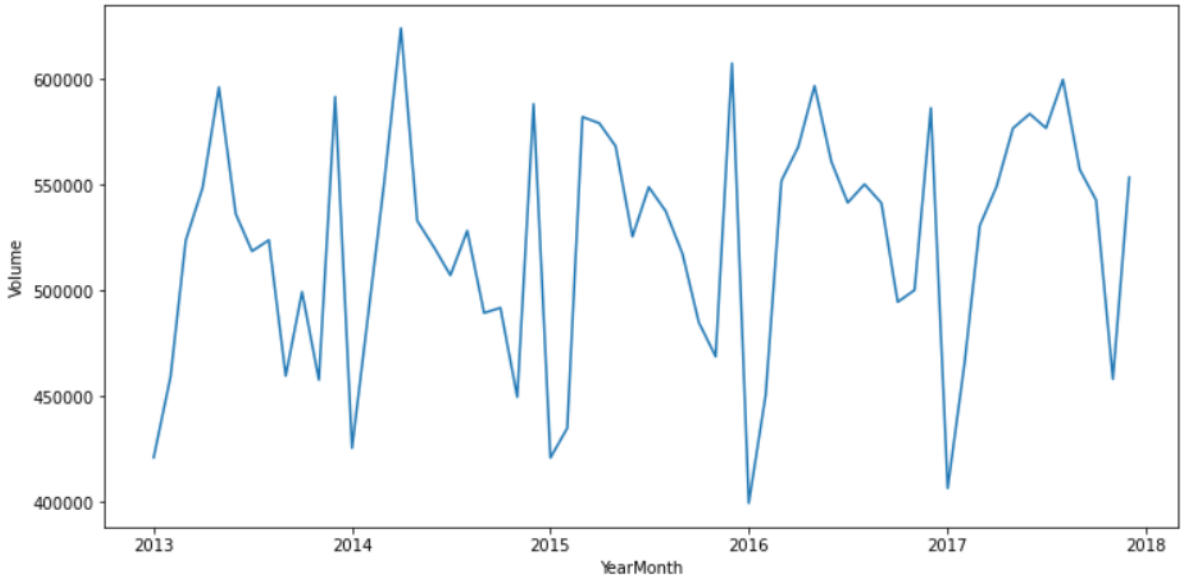

Figura 14 – Evolução do Volume Global com o tempo

Fonte: Elaborada pelo autor

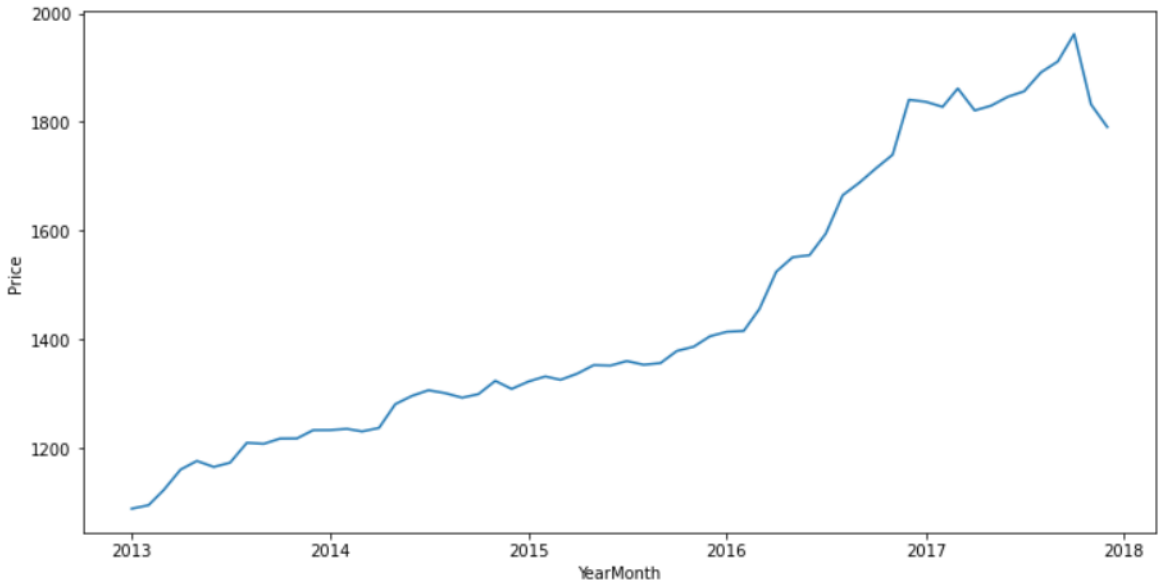

Figura 15 – Evolução do Preço Médio Global com o tempo

Fonte: Elaborada pelo autor

2.5.3 Preço vs Volume

O preço e o volume não demonstram uma associação quando analisados os seus valores globais e mostrado na Figura 16.

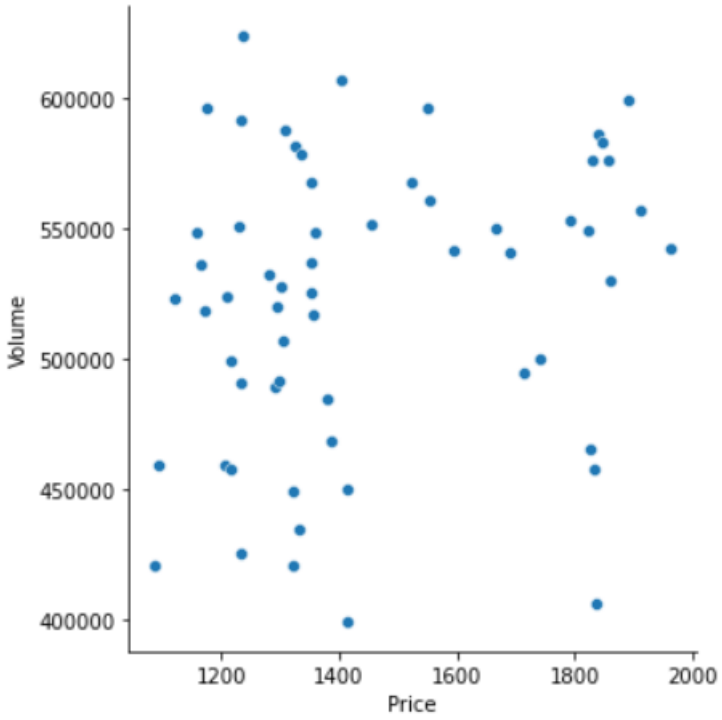

Figura 16 – Gráfico do Preço vs Volume

Fonte: Elaborada pelo autor

2.5.4 Temperatura vs Volume

Há uma clara associação positiva entre a temperatura máxima média e o volume de vendas de cerveja e pode ser verificado na Figura 17.

2.5.5 Litoral vs Volume

Não há uma clara associação entre localização no litoral ou interior e o volume de vendas de cerveja. Figura 18.

2.5.6 População vs Volume

Não parece haver uma associação entre a população da região atendida pelos atacadistas e o volume conforme mostrado na Figura 19.

2.5.7 Renda Familiar vs Volume

Não parece haver uma associação entre a renda familiar e o volume conforme mostrado na Figura 20.

2.5.8 Vendas de Refrigerantes vs Volume

Parece haver uma associação positiva entre a venda de refrigerantes e a venda de cerveja. Vide a Figura 21.

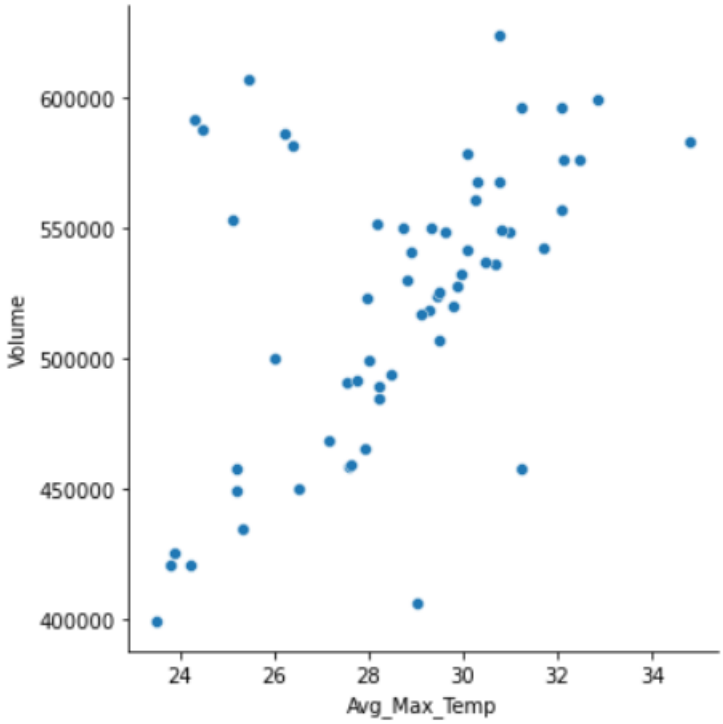

Figura 17 – Gráfico da Temperatura vs Volume

Fonte: Elaborada pelo autor

2.5.9 Produção de Cerveja vs Volume

Parece haver uma associação positiva entre a produção de cerveja da indústria e a venda de cerveja. Vide a Figura 22.

2.5.10 Eventos vs Volume

Não parece haver nenhuma associação entre quaisquer datas festivas ou eventos com Volume, conforme os boxplots da Figura 23.

2.6 Análise Exploratória dos Dados - SKUs de Alto Volume

O objetivo dessa etapa é estudar as diferenças entre as Lojas Atacadistas que vendem as SKUs de maior volume: 1, 2, 3, 4 e 5.

A soma dessas SKUs correpondende a 99,65% do volume total do negócio e o detalhe em hectolitros pode ser visto na Figura 24.

A sobreposição das lojas que comercializam essas SKUs é bastante alta, sendo que a SKU 4 aparece em todas as 58 lojas, a SKU 3 em 57 lojas e as SKUs 1, 2 e 5 em 52 lojas conforme a Tabela 9.

Tabela 9 – Sobreposição das Lojas para as SKUs de Alto Volume

SKU Ag	**1**	**2**	**3**	**4**	**5**	**SKU Ag**	**1**	**2**	**3**	**4**	**5**
1						31					
2						32					
3						33					
4						34					
5						35					
6	X	X	X	X	X	36					
7						37					
8						38					
9						39					
10						40					
11						41					
12						42					
13						43					
14	X	X	X	X	X	44					
15						45					
16						46					
17						47					
18						48					
19						49					
20						50					
21						51					
22						52					
23						53					
24						54					
25						55					
26						56					
27						57					
28						58					
29						59					
30						60					

Fonte: Elaborada pelo autor.

Nota: X: Loja não existe

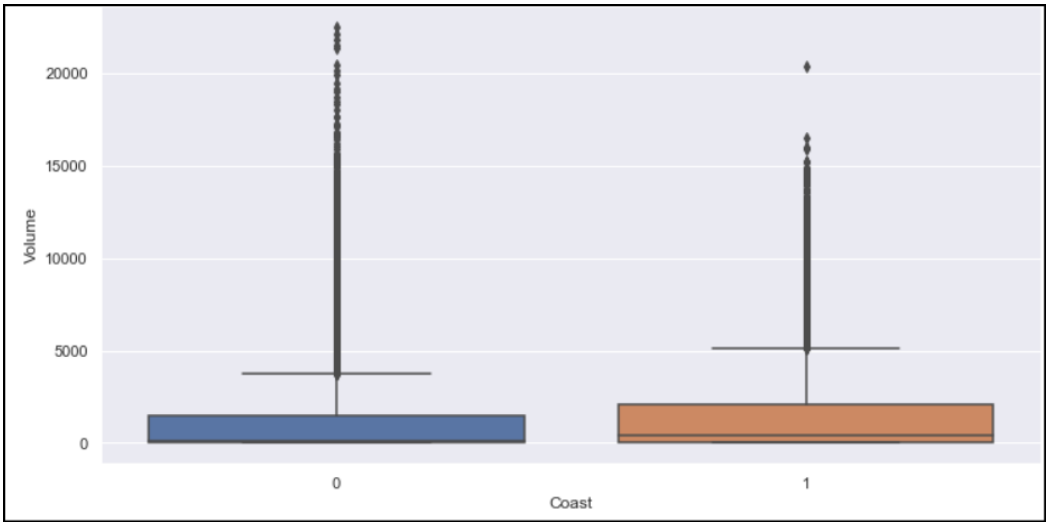

Figura 18 – Gráfico da Localização vs Volume

Fonte: Elaborada pelo autor

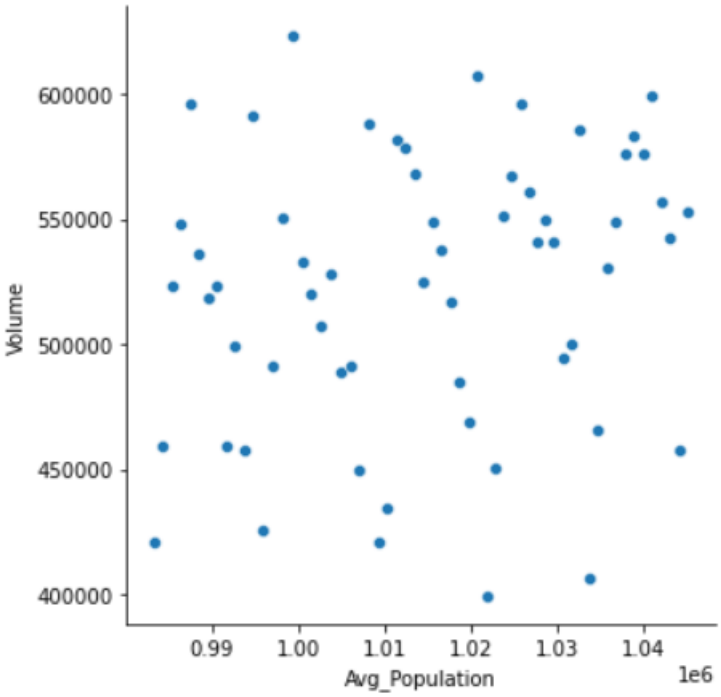

Figura 19 – Gráfico da População vs Volume

Fonte: Elaborada pelo autor

2.7 Análise Exploratória dos Dados - Lojas de Baixo Volume

O objetivo dessa seção é estudar as diferenças entre as lojas atacadistas de baixo volume: 33, 34, 35, 36 e 37.

A soma de seus volumes representa apenas 0,0145% do total e o detalhe em

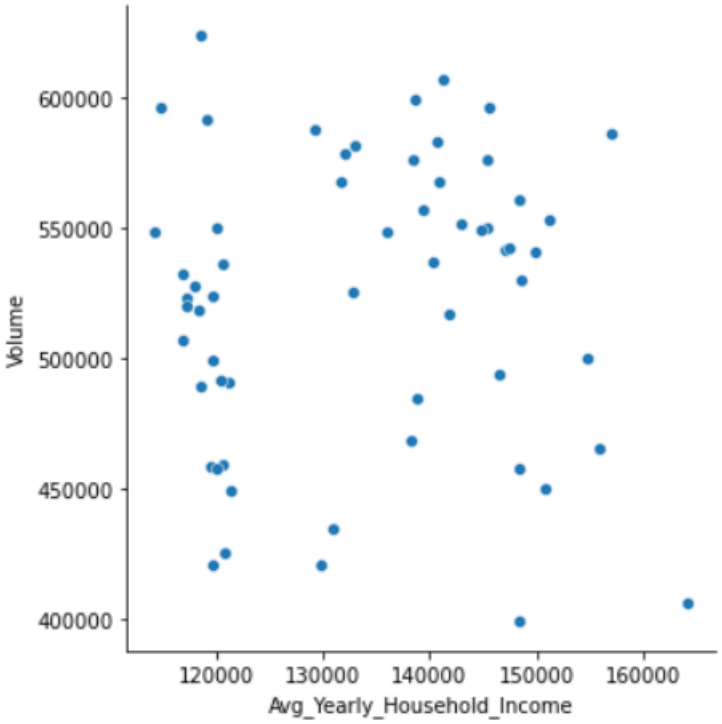

Figura 20 – Gráfico da Renda Familiar vs Volume

Fonte: Elaborada pelo autor

Tabela 10 – SKUs comercializadas pelas Lojas de Pequeno Volume

Agency	**SKU**				
33	3	4	18	X	X
34	3	4	X	22	X
35	3	4	X	X	X
36	3	3	X	22	X
37	X	4	18	X	X

Fonte: Elaborada pelo autor.

hectolitros é mostrado na Figura 25.

Poucas SKUs são comercializadas por essas lojas, conforme a Tabela 10.

Os preços são maiores do que a média e a localização é em regiões mais quentes, de menor população e menor renda e muito provavelmente em regiões do interior do país, devido ao alto desvio-padrão da temperatura. Detalhes na Tabela 11.

2.8 Matriz de Correlação de Pearson

A matriz de correlação de Pearson, para correlações lineares, indica haver uma correlação significativa entre os atributos preditivos *Agency*, *SKU*, *Price*, *Temperature*, *Population*, *Income* e *Coast* com o atributo objetivo *Volume*, conforme a Figura 26.

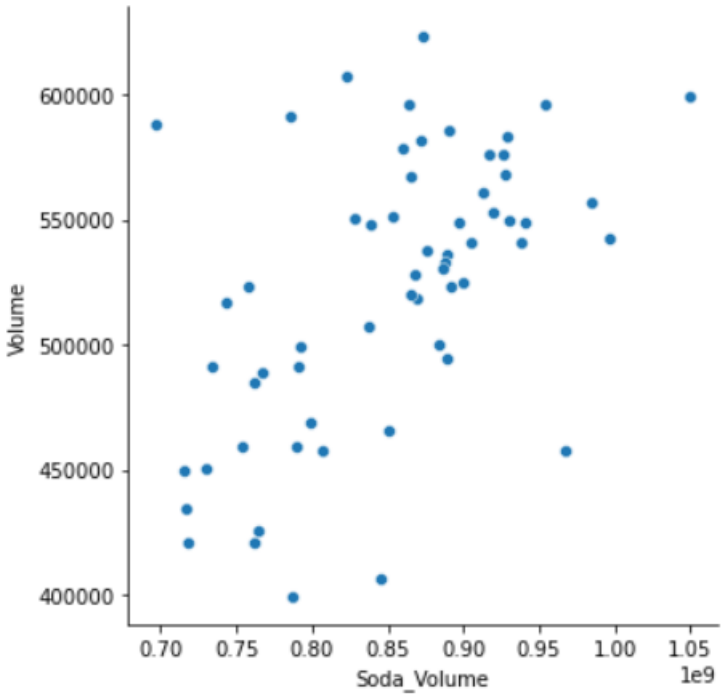

Figura 21 – Gráfico da Venda de Refrigerantes vs Volume

Fonte: Elaborada pelo autor

Tabela 11 – Características da Lojas de Pequeno Volume

Agency	**Price**	**Avg Temp**	**StdDev Temp**	**Population**	**Income**
Global	1545.34	28.61	2.78	1,014,520	133,826
33	1736.02	30.73	3.85	38,557	80,943
34	1901.10	32.18	4.40	11,911	86,122
35	1521.44	28.23	4.16	13,724	82,475
36	1698.49	28.23	4.16	12,726	83,479
37	1929.47	33.81	4.17	31,810	85,714

Fonte: Elaborada pelo autor.

Outras correlações relevantes são identificadas, não envolvendo a variável target:

-Atacadista e temperatura;

-Atacadista e litoral;

-SKU e preço;

-Preço e a venda de refrigerantes e a produção de cerveja;

-Temperatura e a venda de refrigerantes, produção de cerveja e litoral;

-População e renda familiar;

-Vendas de refrigerantes e produção de cerveja.

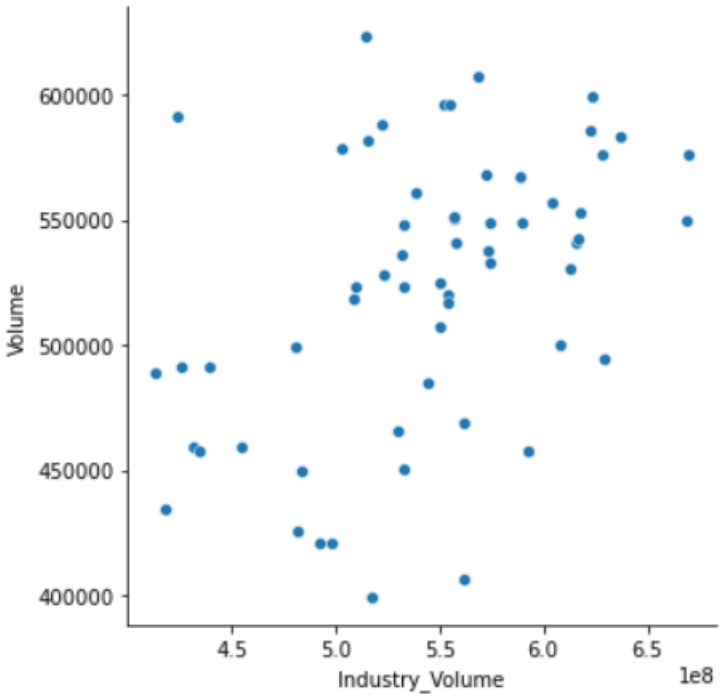

Figura 22 – Gráfico da Produção de Cerveja vs Volume

Fonte: Elaborada pelo autor

2.9 Seleção dos Modelos para Ajuste Fino

Esta etapa consiste em testar diferentes modelos para representar os dados e selecionar os que irão para a fase de ajuste fino.

Serão utilizados os modelos de séries temporais listados na Tabela 12.

Será adotado o modelo *Seasonal Naive* como modelo de base de referência para efeito de comparação de performance dos demais modelos preditivos.

Foi feita uma prospecção à fim de conhecer-se os modelos que mais se adequam às mais diferentes combinações de atributo, conforme a Tabela 13

A Tabela 14 apresenta os modelos que tiveram melhor desempenho segundo a combinação de valores extremos dos atributos preditivos.

A Tabela 15 apresenta as melhores métricas obtidas pelos modelos que tiveram melhor desempenho segundo a combinação de atributos.

2.10 Agrupamento por Loja e Produto

Em princípio, temos que criar modelos preditivos para cada combinação de Loja Atacadista (*Agency*) e Produto/SKU perfazendo um total de 350 modelos preditivos. O detalhe das combinações é apresentado na Tabela 16.

Tabela 12 – Tabela de Modelos de Séries Temporais

Número	**Modelo**	**Obs**
1	Naive	Última observação
2	Seasonal Naive	Última observação sazonal
3	Holt-Winters Additive-Additive	Tendência-Sazonalidade
4	Holt-Winters Additive-Multiplicative	Tendência-Sazonalidade
5	Holt-Winters Multiplicative-Additive	Tendência-Sazonalidade
6	Holt-Winters Multiplicative-Multiplicative	Tendência-Sazonalidade
7	Sarima	
8	Sarimax	2 Variáveis Exógenas
9	Theta	
10	LSTM_11	No. Camadas LSTM e Dense
11	LSTM_21	No. Camadas LSTM e Dense

Fonte: Elaborada pelo autor.

Tabela 13 – Tabela de Combinação de Valores Extremos dos Atributos

Código	**Estratificação**
HPOP-HVOL	Alta População - Alto Volume
HPOP-LVOL	Alta População - Baixo Volume
LPOP-HVOL	Baixa População - Alto Volume
LPOP-LVOL	Baixa População - Baixo Volume
HINC-HVOL	Alta Renda - Alto Volume
HINC-LVOL	Alta Renda - Baixo Volume
LINC-HVOL	Baixa Renda - Alto Volume
LINC-LVOL	Baixa Renda - Baixo Volume
HTEMP-HVOL	Alta Temperatura - Alto Volume
HTEMP-LVOL	Alta Temperatura - Baixo Volume
LTEMP-HVOL	Baixa Temperatura - Alto Volume
LTEMP-LVOL	Baixa Temperatura - Baixo Volume
HDVTEMP-HVOL	Alto Desvio-Padrão Temperatura - Alto Volume
HDVTEMP-LVOL	Alto Desvio-Padrão Temperatura - Baixo Volume
LDVTEMP-HVOL	Baixo Desvio-Padrão Temperatura - Alto Volume
LDVTEMP-LVOL	Baixo Desvio-Padrão Temperatura - Baixo Volume

Fonte: Elaborada pelo autor.

Nota: H-High/alto; L-Low/baixo; INC-Income/renda

Tabela 14 – Tabela dos Melhores Modelos segundo a Combinação de Atributos

Código	**Agency**	**SKU**	**MAPE**	**RMSE**	**TU**	**POCID**
HPOP-HVOL	2	3	4,6	3,5	3,4,5,6	3,4,5,6
HPOP-LVOL	2	11	10	1	1,10,11	10
LPOP-HVOL	34	4	10	10	10	3,9,10,11
LPOP-LVOL	34	22	10	1,2,7,8,10,11	10	11
HINC-HVOL	30	3	2	2	2	2,3,4,7,8
HINC-LVOL	30	6	6	1,3,4,6,7,8,9,10,11	6	3,6,7,9
LINC-HVOL	23	5	8	8	8	3,4,5,6,7,8
LINC-LVOL	23	21	11	10,11	11	10,11
HTEMP-HVOL	37	4	11	9,10	9	10
HTEMP-LVOL	37	18	11	1,2,3,4,5,6,7,8,10,11	6	11
LTEMP-HVOL	60	1	2,7	2,7	2,7	3,4,5,6
LTEMP-LVOL	60	23	8	8,10	10	8,11
HDVTEMP-HVOL	26	11	2	1,2,7,11	2	3
HDVTEMP-LVOL	26	5	10	1,2,3,7,8,9,10,11	3	7,8
LDVTEMP-HVOL	9	3	2	2	2	2,9,10,11
LDVTEMP-LVOL	9	21	10	10	10	3,9,10,11

Fonte: Elaborada pelo autor.

Tabela 15 – Tabela das Melhores Métricas segundo a Combinação de Atributos

Código	**Agency**	**SKU**	**MAPE**	**RMSE**	**TU**	**POCID**
HPOP-HVOL	2	3	8,22	1721	0,38	82,0
HPOP-LVOL	2	11	51,57	3	0,00	64,0
LPOP-HVOL	34	4	72,66	14	0,59	55,0
LPOP-LVOL	34	22	4,4e+17	2	0,56	64,0
HINC-HVOL	30	3	7,13	1098	1,00	100,0
HINC-LVOL	30	6	77,88	1	0,26	64,0
LINC-HVOL	23	5	6,83	57	0,25	82,0
LINC-LVOL	23	21	54,97	0	0,31	55,0
HTEMP-HVOL	37	4	53,30	33	0,81	55,0
HTEMP-LVOL	37	18	288,52	1	0,53	73,0
LTEMP-HVOL	60	1	10,43	1425	1,00	73,0
LTEMP-LVOL	60	23	4,5e+16	0	0,29	64,0
HDVTEMP-HVOL	26	11	22,26	0	1,00	82,0
HDVTEMP-LVOL	26	5	59,66	0	0,65	73,0
LDVTEMP-HVOL	9	3	13,32	1169	1,00	73,0
LDVTEMP-LVOL	9	21	78,08	0	0,73	55,0

Fonte: Elaborada pelo autor.

Tabela 16 – Combinação de Atacadistas (*Agency*) e SKU

Ag	**SKU**	#	**Ag**	**SKU**	#
1	1,2,3,4,5,11	6	31	1,2,3,4,5,8	6
2	1,2,3,4,5,11,12,31,34	9	32	1,2,3,4,5,14	6
3	1,2,3,4,5,32	6	33	3,4,18	3
4	1,2,3,4,5	5	34	3,4,22	3
5	1,2,3,4,5,14,21,23,26	9	35	3,4,32	3
6	NA	NA	36	3,4,22	3
7	1,2,3,4,5,21	6	37	4,18	2
8	1,2,3,4,5	5	38	1,2,3,4,5,7,14,21	8
9	1,2,3,4,5,21,27	7	39	1,2,3,4,5,7	6
10	1,2,3,4,5,21,23,32	8	40	3,4,18	3
11	1,2,3,4,5	5	41	1,2,3,4,5	5
12	1,2,3,4,5,7	6	42	1,2,3,4,5	5
13	1,2,3,4,5,7,20	7	43	1,2,3,4,5	5
14	NA	NA	44	1,2,3,4,5	5
15	1,2,3,4,5	5	45	1,2,3,4,5,8	6
16	1,2,3,4,5,7,20	7	46	1,2,3,4,5,17	6
17	1,2,3,4,5	5	47	1,2,3,4,5,17	6
18	1,2,3,4,5	5	48	1,2,3,4,5,7,17,23,28	9
19	1,2,3,4,5,8	6	49	1,2,3,4,5,7,23,34	8
20	1,2,3,4,5,7,21,23	8	50	1,2,3,4,5,17	6
21	1,2,3,4,5,21	6	51	1,2,3,4,5,7,17	7
22	1,2,3,4,5	5	52	1,2,3,4,5	5
23	1,2,3,4,5,21	6	53	1,2,3,4,5,7,15,24	8
24	1,2,3,4,5,21	6	54	1,2,3,4,5,14,15	7
25	1,2,3,4,5,21	6	55	1,2,3,4,5	5
26	1,2,3,4,5,11,18	7	56	1,2,3,4,5	5
27	1,2,3,4,5,6,8	7	57	1,2,3,4,5,7,17,23	8
28	1,2,3,4,5,6,8	7	58	1,2,3,4,5,7,17,23	8
29	1,2,3,4,5,6,8	7	59	1,2,3,4,5,7,17	7
30	1,2,3,4,5,6,8	7	60	1,2,3,4,5,7,23	7

Fonte: Elaborada pelo autor.

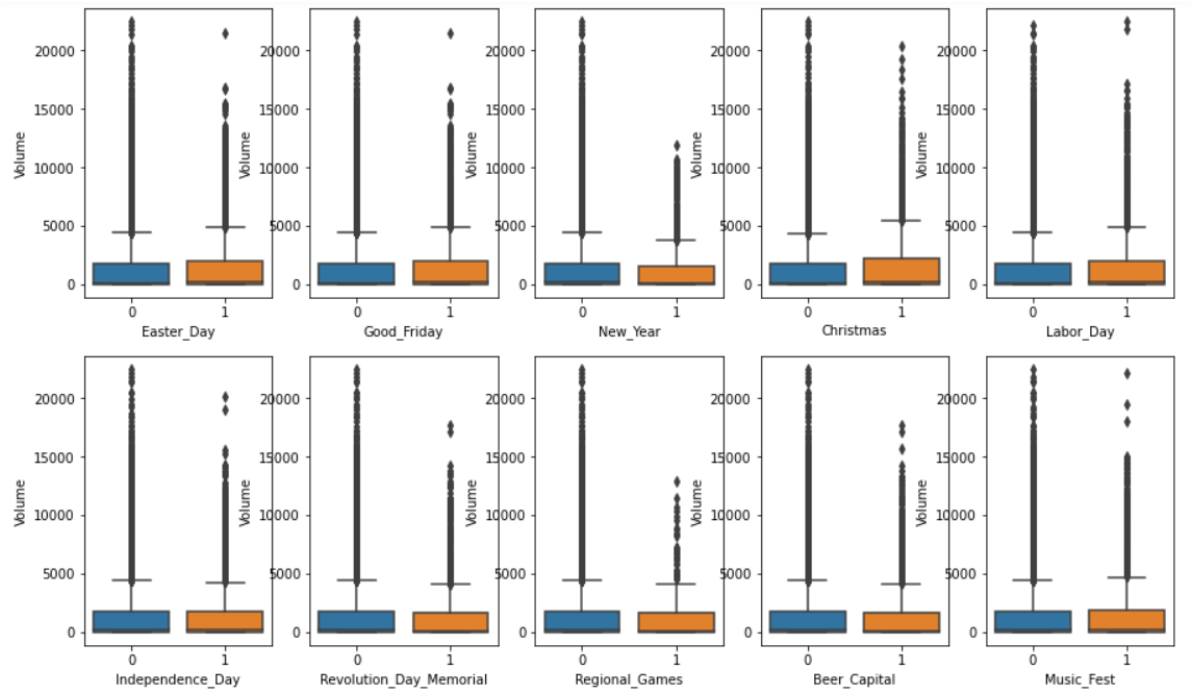

Figura 23 – Boxplots dos Eventos vs Volume

Fonte: Elaborada pelo autor

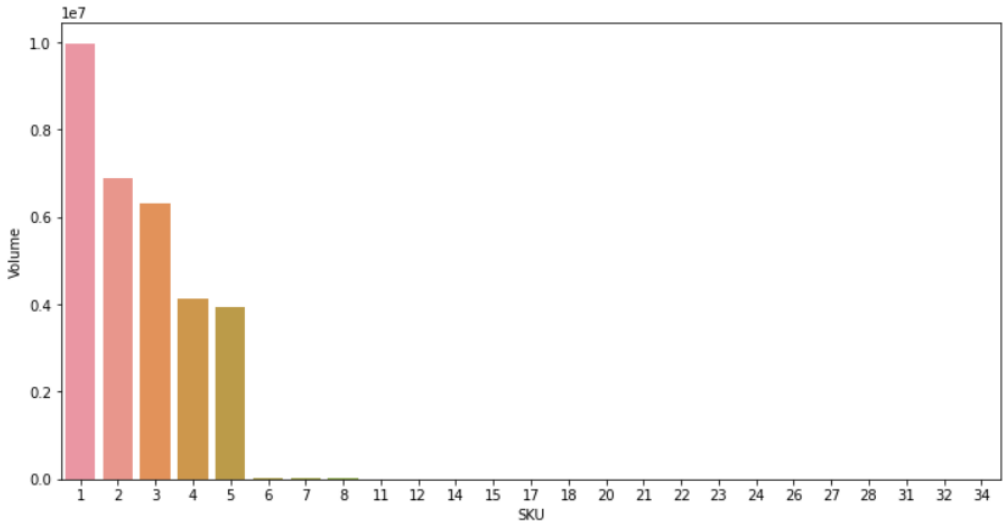

Figura 24 – Volume por SKU

Fonte: Elaborada pelo autor

À fim de evitar essa complexidade e reduzir a quantidade de modelos, foi realizada uma clusterização por Loja e Produto, usando a técnica K-médias.

Para a correta aplicação da técnica, as variáveis Preço, Temperatura, População, Renda e Volume foram normalizadas com MinMaxScaler e StandardScaler da biblioteca *scikit-learn*, as variáveis SKU foram dummiezadas com *one-hot encoding*, enquanto que

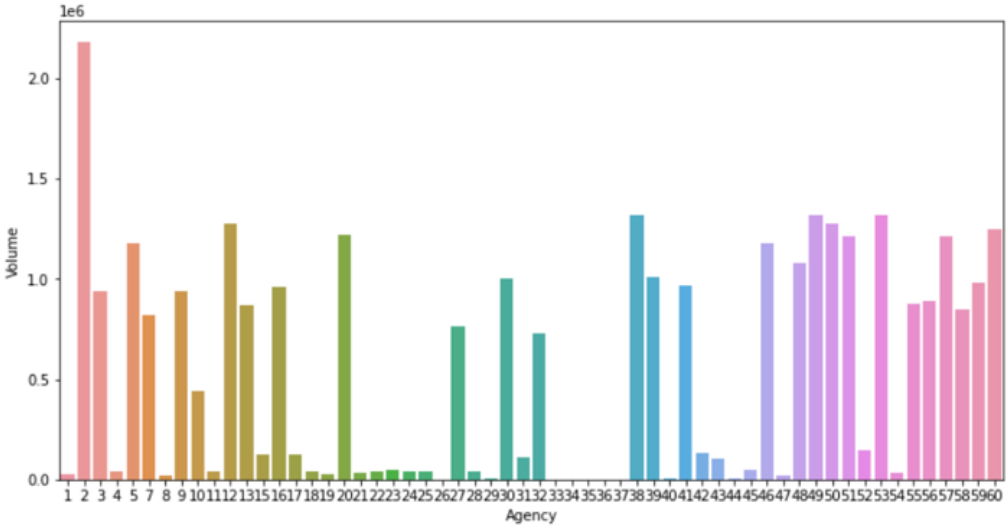

Figura 25 – Volume por Loja Atacadista

Fonte: Elaborada pelo autor

as variáveis de datas festivas e feriados foram desprezadas por não trazerem informação diferencial.

Foram utilizadas 3 diferentes métricas: NMI (*Normalized Mutual information*), o método do cotovelo e a *Silhouette score*.

O agrupamento utilizou apenas os dados de treinamento (os 4 primeiros anos), e nesse caso as (Agency,SKU) = (10,21) e (48,28) não estavam presentes e não foram consideradas nem tratadas nesse estudo, gerando um total de 348 combinações Loja x Produto em vez das 350 iniciais.

O resultado que apresentou a distinção mais clara entre os grupos foi a normalização MinMaxScaler e o método do cotovelo e está representado pela Figura 27, de onde obtemos que o melhor k=6.

O modelo gerou para cada combinação Loja x Produto um grupo para cada mês, podendo para uma mesma combinação definir grupos diferentes dependendo do período analisado.

A definição final do grupo foi feita usando-se a moda, de tal modo que cada combinação Loja x Produto ficou atribuída a apenas um grupo. Após aplicação da moda, verificou-se que dos 6 grupos gerados pelo k-médias restaram apenas 5, ou seja um grupo foi eliminado após aplicação dessa técnica.

O resultado final dos grupos foi rotulado como A, B, C, D e E e está apresentado na Tabela 17, Tabela 18, Tabela 19, Tabela 20 e Tabela 21.

Tabela 17 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 1 a 7

SKU Ag	01	02	03	04	05	06	07
1	A	A	A	A	A	-	-
2	E	E	E	C	C	-	-
3	C	C	E	C	C	-	-
4	B	B	B	B	B	-	-
5	E	C	C	C	C	-	-
6	-	-	-	-	-	-	-
7	E	C	C	C	C	-	-
8	B	B	B	B	B	-	-
9	C	C	E	C	C	-	-
10	D	D	D	D	D	-	-
11	B	B	B	B	B	-	-
12	E	C	C	C	C	-	C
13	E	D	D	D	D	-	D
14	-	-	-	-	-	-	-
15	A	A	A	A	A	-	-
16	E	C	C	C	C	-	C
17	A	A	A	A	A	-	-
18	B	B	B	B	B	-	-
19	B	B	B	B	B	-	-
20	D	E	D	D	D	-	D
21	B	B	B	B	B	-	-
22	B	B	B	B	B	-	-
23	B	B	B	B	B	-	-
24	B	B	B	B	B	-	-
25	B	B	B	B	B	-	-
26	B	B	B	B	B	-	-
27	C	C	C	C	C	C	-
28	B	B	B	B	B	B	-
29	B	B	B	B	B	B	-
30	C	C	E	C	C	C	-

SKU Ag	01	02	03	04	05	06	07
31	B	B	B	B	B	-	-
32	D	D	D	D	D	-	-
33	-	-	B	B	-	-	-
34	-	-	B	B	-	-	-
35	-	-	A	A	-	-	-
36	-	-	A	A	-	-	-
37	-	-	-	B	-	-	-
38	E	D	D	D	D	-	D
39	D	D	D	D	D	-	D
40	-	-	B	A	-	-	-
41	D	D	D	D	D	-	-
42	B	B	B	B	B	-	-
43	A	A	A	A	A	-	-
44	A	A	A	A	A	-	-
45	B	B	B	B	B	-	-
46	C	C	C	C	C	-	-
47	B	B	B	B	B	-	-
48	C	C	E	C	C	-	C
49	D	D	E	D	D	-	D
50	D	E	D	D	D	-	D
51	C	C	C	C	C	-	C
52	B	B	B	B	B	-	-
53	D	E	D	D	D	-	D
54	A	A	A	A	A	-	-
55	C	C	C	C	C	-	-
56	C	C	E	C	C	-	-
57	E	D	C	C	C	-	C
58	D	D	D	D	D	-	D
59	E	C	C	C	C	-	C
60	E	C	C	C	C	-	C

Fonte: Elaborada pelo autor.

Tabela 18 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 8 a 14

SKU Ag	08	09	10	11	12	13	14
1	-	-	-	A	-	-	-
2	-	-	-	C	C	-	-
3	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-
5	-	-	-	-	-	-	C
6	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-
13	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-
19	B	-	-	-	-	-	-
20	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-
26	-	-	-	B	-	-	-
27	C	-	-	-	-	-	-
28	B	-	-	-	-	-	-
29	B	-	-	-	-	-	-
30	C	-	-	-	-	-	-

SKU Ag	08	09	10	11	12	13	14
31	B	-	-	-	-	-	-
32	-	-	-	-	-	-	D
33	-	-	-	-	-	-	-
34	-	-	-	-	-	-	-
35	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-
37	-	-	-	-	-	-	-
38	-	-	-	-	-	-	C
39	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-
45	B	-	-	-	-	-	-
46	-	-	-	-	-	-	-
47	-	-	-	-	-	-	-
48	-	-	-	-	-	-	-
49	-	-	-	-	-	-	-
50	-	-	-	-	-	-	-
51	-	-	-	-	-	-	-
52	-	-	-	-	-	-	-
53	-	-	-	-	-	-	-
54	-	-	-	-	-	-	A
55	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-
57	-	-	-	-	-	-	-
58	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-
60	-	-	-	-	-	-	-

Fonte: Elaborada pelo autor.

Tabela 19 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 15 a 21

SKU Ag	15	16	17	18	19	20	21
1	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-
5	-	-	-	-	-	-	C
6	-	-	-	-	-	-	-
7	-	-	-	-	-	-	C
8	-	-	-	-	-	-	-
9	-	-	-	-	-	-	C
10	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-
13	-	-	-	-	-	D	-
14	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-
16	-	-	-	-	-	C	-
17	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-
20	-	-	-	-	-	-	D
21	-	-	-	-	-	-	A
22	-	-	-	-	-	-	-
23	-	-	-	-	-	-	A
24	-	-	-	-	-	-	A
25	-	-	-	-	-	-	A
26	-	-	-	B	-	-	-
27	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-

SKU Ag	15	16	17	18	19	20	21
31	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-
33	-	-	-	B	-	-	-
34	-	-	-	-	-	-	-
35	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-
37	-	-	-	B	-	-	-
38	-	-	-	-	-	-	C
39	-	-	-	-	-	-	-
40	-	-	-	B	-	-	-
41	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-
45	-	-	-	-	-	-	-
46	-	-	C	-	-	-	-
47	-	-	B	-	-	-	-
48	-	-	C	-	-	-	-
49	-	-	-	-	-	-	-
50	-	-	C	-	-	-	-
51	-	-	C	-	-	-	-
52	-	-	-	-	-	-	-
53	D	-	-	-	-	-	-
54	A	-	-	-	-	-	-
55	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-
57	-	-	C	-	-	-	-
58	-	-	D	-	-	-	-
59	-	-	C	-	-	-	-
60	-	-	-	-	-	-	-

Fonte: Elaborada pelo autor.

Tabela 20 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 22 a 28

SKU Ag	**22**	**23**	**24**	**25**	**26**	**27**	**28**
1	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-
5	-	C	-	-	C	-	-
6	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-
9	-	-	-	-	-	C	-
10	-	D	-	-	-	-	-
11	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-
13	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-
20	-	D	-	-	-	-	-
21	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-

SKU Ag	**22**	**23**	**24**	**25**	**26**	**27**	**28**
31	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-
33	-	-	-	-	-	-	-
34	B	-	-	-	-	-	-
35	-	-	-	-	-	-	-
36	A	-	-	-	-	-	-
37	-	-	-	-	-	-	-
38	-	-	-	-	-	-	-
39	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-
45	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-
47	-	-	-	-	-	-	-
48	-	C	-	-	-	-	-
49	-	C	-	-	-	-	-
50	-	-	-	-	-	-	-
51	-	-	-	-	-	-	-
52	-	-	-	-	-	-	-
53	-	-	D	-	-	-	-
54	-	-	-	-	-	-	-
55	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-
57	-	C	-	-	-	-	-
58	-	D	-	-	-	-	-
59	-	-	-	-	-	-	-
60	-	C	-	-	-	-	-

Fonte: Elaborada pelo autor.

Tabela 21 – Agrupamento de Lojas Atacadistas (*Agency*) para as SKUs 29 a 34

SKU Ag	**29**	**30**	**31**	**32**	**33**	**34**
1	-	-	-	-	-	-
2	-	-	C	-	-	C
3	-	-	-	C	-	-
4	-	-	-	-	-	-
5	-	-	-	-	-	-
6	-	-	-	-	-	-
7	-	-	-	-	-	-
8	-	-	-	-	-	-
9	-	-	-	-	-	-
10	-	-	-	D	-	-
11	-	-	-	-	-	-
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	-	-	-	-	-	-
15	-	-	-	-	-	-
16	-	-	-	-	-	-
17	-	-	-	-	-	-
18	-	-	-	-	-	-
19	-	-	-	-	-	-
20	-	-	-	-	-	-
21	-	-	-	-	-	-
22	-	-	-	-	-	-
23	-	-	-	-	-	-
24	-	-	-	-	-	-
25	-	-	-	-	-	-
26	-	-	-	-	-	-
27	-	-	-	-	-	-
28	-	-	-	-	-	-
29	-	-	-	-	-	-
30	-	-	-	-	-	-

SKU Ag	**29**	**30**	**31**	**32**	**33**	**34**
31	-	-	-	-	-	-
32	-	-	-	-	-	-
33	-	-	-	-	-	-
34	-	-	-	-	-	-
35	-	-	-	A	-	-
36	-	-	-	-	-	-
37	-	-	-	-	-	-
38	-	-	-	-	-	-
39	-	-	-	-	-	-
40	-	-	-	-	-	-
41	-	-	-	-	-	-
42	-	-	-	-	-	-
43	-	-	-	-	-	-
44	-	-	-	-	-	-
45	-	-	-	-	-	-
46	-	-	-	-	-	-
47	-	-	-	-	-	-
48	-	-	-	-	-	-
49	-	-	-	-	-	C
50	-	-	-	-	-	-
51	-	-	-	-	-	-
52	-	-	-	-	-	-
53	-	-	-	-	-	-
54	-	-	-	-	-	-
55	-	-	-	-	-	-
56	-	-	-	-	-	-
57	-	-	-	-	-	-
58	-	-	-	-	-	-
59	-	-	-	-	-	-
60	-	-	-	-	-	-

Fonte: Elaborada pelo autor.

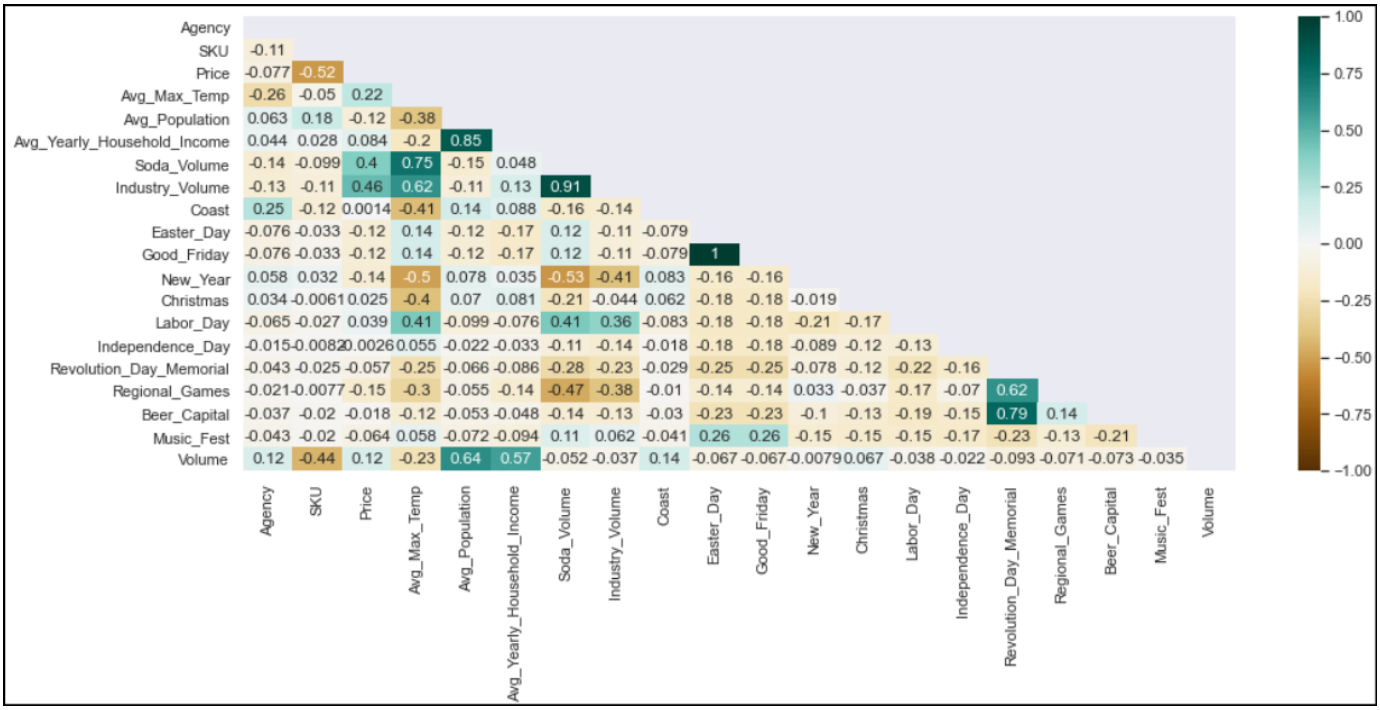

Figura 26 – Matriz de Correlação de Pearson

Fonte: Elaborada pelo autor

2.11 Ajuste fino dos Modelos Finais

Foi realizado um ajuste fino para cada um dos 5 grupos de Loja x Produto obtido na etapa anterior.

Para essa etapa foram mantidos os modelos *Naive* e *Seasonal Naive* como *baselines* pela sua simplicidade, rapidez de processamento e baixo consumo de recursos computacionais.

Ainda foram mantidos os métodos SARIMA e LSTM pela sua possibilidade de otimização e capacidade de predição de resultados. Para ambos, foi feito o ajuste fino considerando os modelos Univariados e Multivariados.

Para o SARIMA Univariado foi feita a otimização dos hiperparâmetros max_iter e $info\_crit$, que representam o número máximo de avaliações da função e o critério de seleção usado para o melhor modelo ARIMA, respectivamente, enquanto que para o SARIMA Multivariado a otimização foi na quantidade e em quais variáveis exógenas a serem utilizadas na modelagem.

Para o LSTM Univariado e Multivariado otimizou-se *batch_size* e *timesteps* do gerador de *batches*, que correspondem ao tamanho do *batch* e o número de exemplos a serem considerados para a previsão futura.

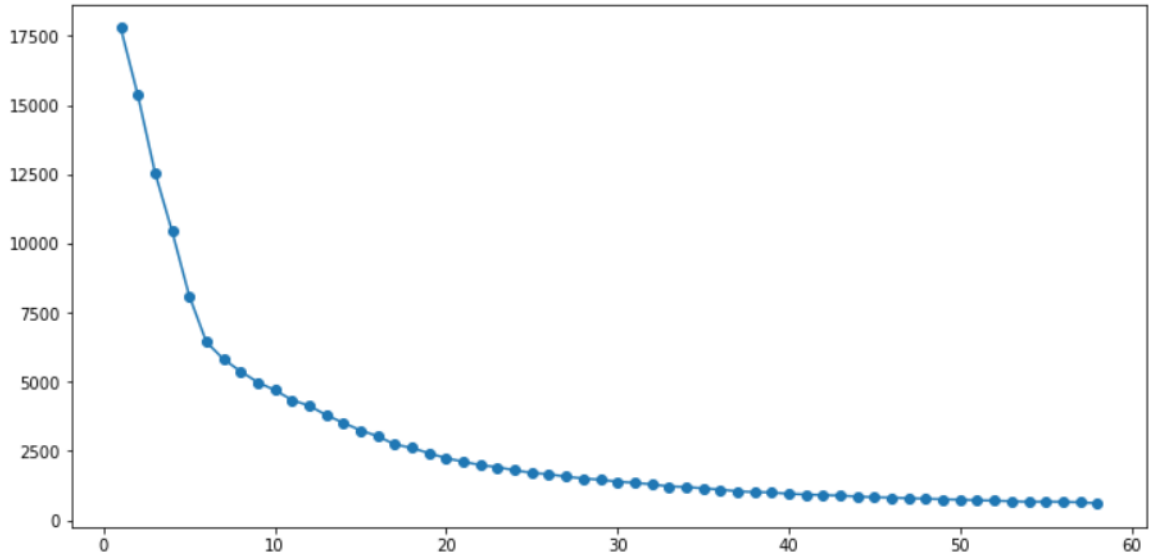

Figura 27 – Gráfico do Método do Cotovelo para Agrupamento Loja-Produto

Fonte: Elaborada pelo autor

3 RESULTADOS

3.1 Perfil dos Grupos

Os grupos A e B representam os baixos volumes, o C o volume médio enquanto que D e E os volumes altos. O resumo do resultado dessa e das demais características é apresentada na Tabela 22.

O o comparativo detalhado dos grupos é mostrado na Figura 28, Figura 29, Figura 30, Figura 31, Figura 32 e Figura 33.

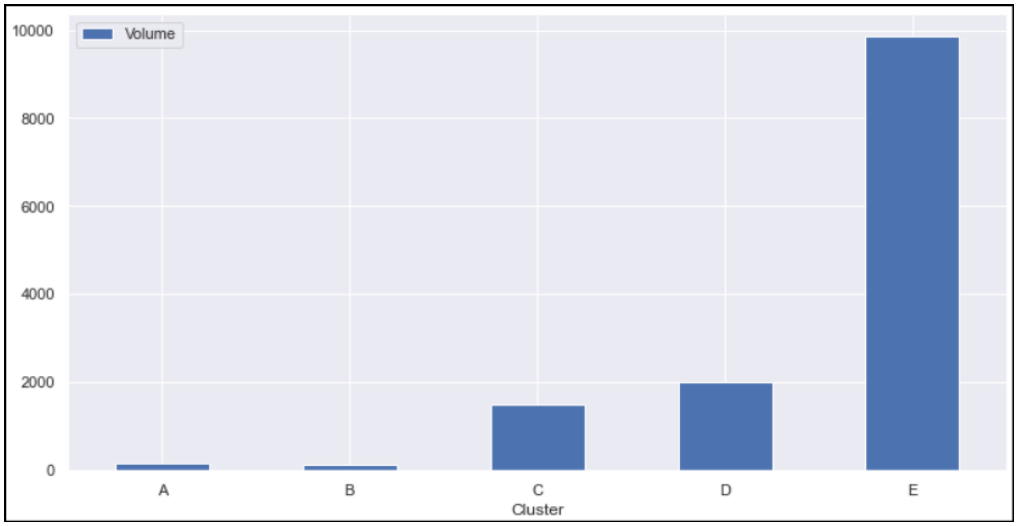

Figura 28 – Comparativo do Volume por Grupo

Fonte: Elaborada pelo autor

Tabela 22 – Resumo do Perfil dos Grupos

Grupo / Característica	A	B	C	D	E
Volume	Baixo	Baixo	Médio	Alto	Alto
Preço	Baixo	Alto	Médio	Alto	Médio
Temperatura	Baixa	Alta	Média	Baixa	Média
População	Baixa	Baixa	Alta	Média	Alta
Renda	Baixa	Baixa	Alta	Média	Alta
Proximidade Litoral	Baixa	Baixa	Média	Alta	Alta

Fonte: Elaborada pelo autor.

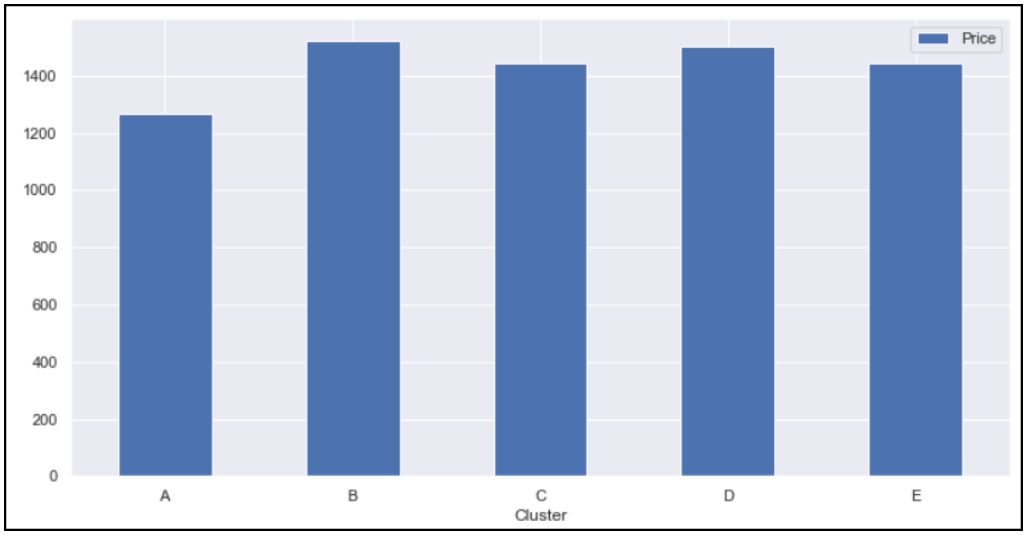

Figura 29 – Comparativo do Preço por Grupo

Fonte: Elaborada pelo autor

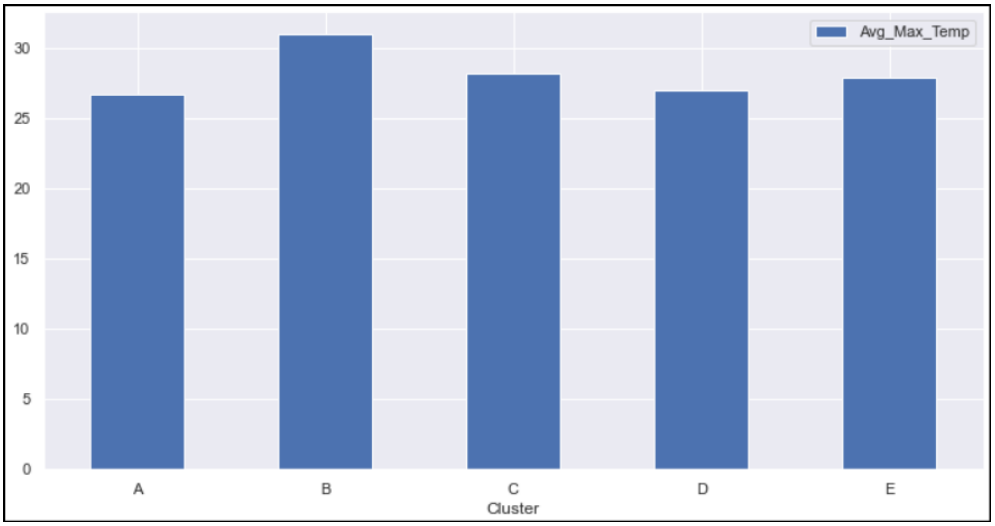

Figura 30 – Comparativo da Temperatura da Região por Grupo

Fonte: Elaborada pelo autor

3.2 Resumo dos Modelos por Grupo

O resultados do melhor modelo para cada grupo, bem como o exemplar de Loja e Produto utilizado estão apresentados na Tabela 23.

Obteve-se que modelos simples como o *Naive* e *Seasonal Naive* tiveram desempenho similar em alguns casos aos modelos mais sofisticados.

O *Seasonal Naive* foi melhor em um grupo, o LSTM também em um e o SARIMA em três situações.

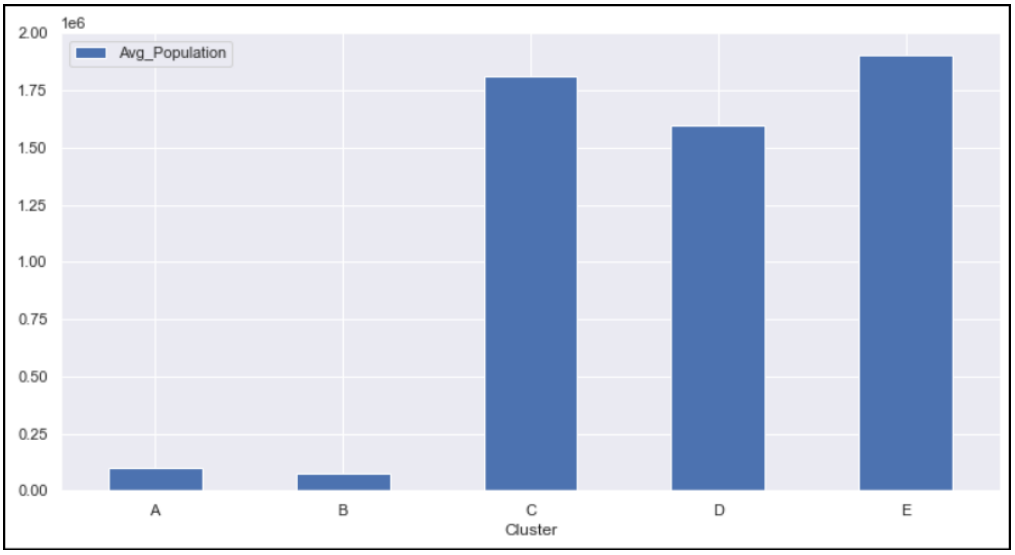

Figura 31 – Comparativo da População da Região por Grupo

Fonte: Elaborada pelo autor

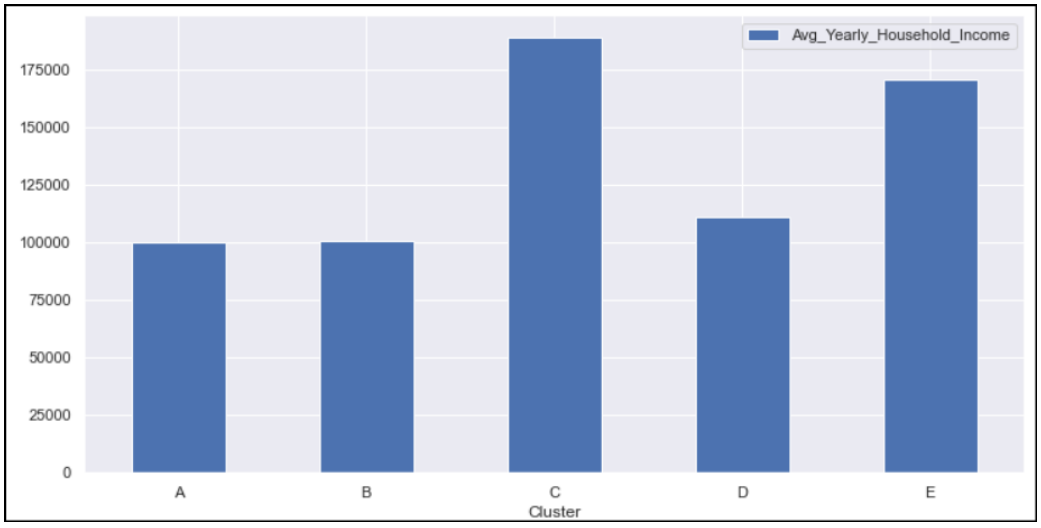

Figura 32 – Comparativo da Renda da População da Região por Grupo

Fonte: Elaborada pelo autor

A magnitude do RMSE depende do valor da variável sendo predita e não deve ser comparada entre um grupo e outro.

O erro medido pelo MAPE tem a tendência de ser maior para as menores magnitudes das variáveis a serem preditas, não sendo possível seu cálculo para valores zero.

3.3 Resultados do Grupo A

A série temporal do Grupo A pode ser vista na Figura 34 e os resultados na Tabela 24 e Tabela 25.

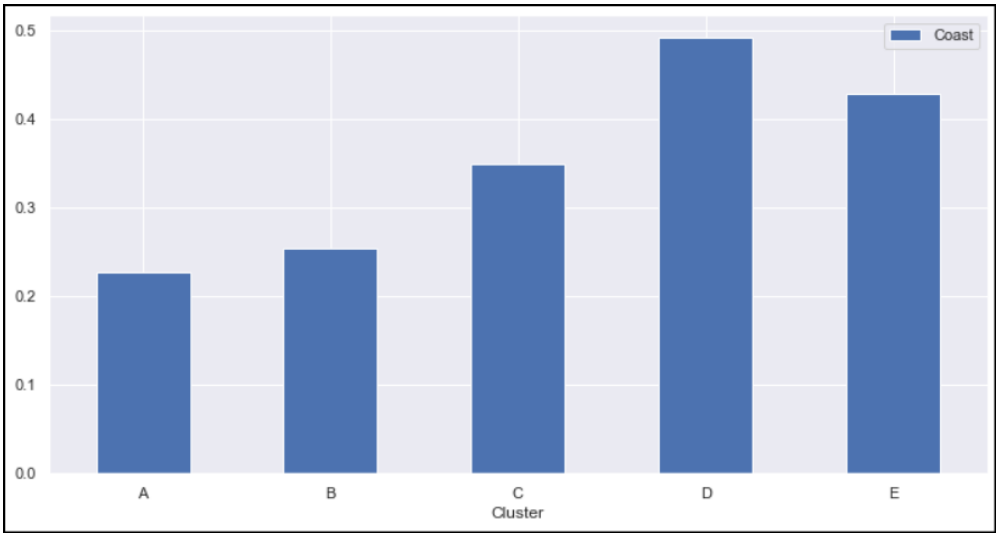

Figura 33 – Comparativo da Proximidade com o Litoral da Região por Grupo

Fonte: Elaborada pelo autor

Tabela 23 – Melhor Modelo de Previsão por Grupo

Cluster	Loja	Produto	Melhor Modelo	RMSE (hl)	MAPE (%)
A	23	21	LSTM Univariado	0,91	74,30
B	26	11	Seasonal Naive	0,76	22,26
C	60	23	SARIMA Multivariado	0,80	-
D	32	2	SARIMA Univariado	388,79	11,79
E	2	3	SARIMA Multivariado	2217,97	11,10

Fonte: Elaborada pelo autor.

Existem muitos valores iguais a zero no período inicial que não foram eliminados pois a intenção é uma generalização do modelo para todas as combinações possíveis de Loja x Produto dentro do mesmo grupo.

Um impacto dessa decisão nesse exemplo selecionado foi que haviam muito poucos dados de treinamento, o que impossibilitou a construção dos modelos SARIMA.

Como os dados de teste eram muito diferentes dos dados de treinamento, os modelos *Naive* tiveram mau desempenho, e os modelos LSTM forem melhores.

3.4 Resultados do Grupo B

A série temporal do Grupo B pode ser vista na Figura 35 e os resultados na Tabela 26 e Tabela 27.

O exemplo modelado possui o conjunto de teste com comportamento similar ao período imediatamente anterior, o que favoreceu os modelos *Naive*.

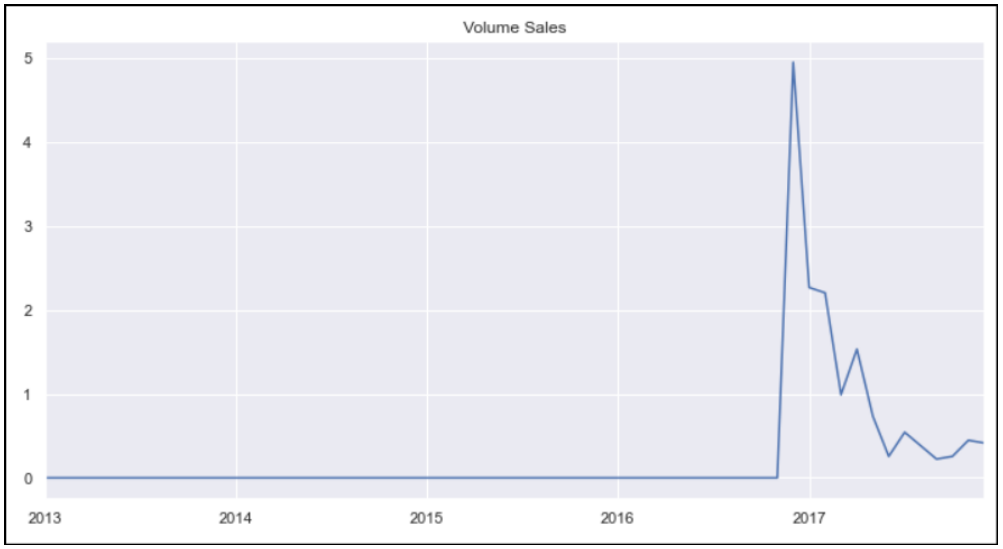

Figura 34 – Série Temporal Grupo A - Loja 23 - Produto 21

Fonte: Elaborada pelo autor

Tabela 24 – Grupo A - Previsões por Modelo

Mês	Real (hl)	Naive	Seasonal Naive	SARIMA uni	SARIMA multi	LSTM uni	LSTM multi
2017-01-01	2.268	4.952	0.000	NaN	NaN	0.309	1.179
2017-02-01	2.205	4.952	0.000	NaN	NaN	0.389	1.492
2017-03-01	0.990	4.952	0.000	NaN	NaN	0.417	1.617
2017-04-01	1.534	4.952	0.000	NaN	NaN	0.262	1.723
2017-05-01	0.735	4.952	0.000	NaN	NaN	0.237	1.665
2017-06-01	0.256	4.952	0.000	NaN	NaN	0.170	1.669
2017-07-01	0.543	4.952	0.000	NaN	NaN	0.112	1.646
2017-08-01	0.383	4.952	0.000	NaN	NaN	0.082	1.645
2017-09-01	0.224	4.952	0.000	NaN	NaN	0.068	1.641
2017-10-01	0.256	4.952	0.000	NaN	NaN	0.058	1.667
2017-11-01	0.447	4.952	0.000	NaN	NaN	0.046	1.721
2017-12-01	0.415	4.952	4.952	NaN	NaN	0.057	1.785

Fonte: Elaborada pelo autor.

O histórico do período total de treinamento com comportamento diferente do período de teste fez com que os modelos SARIMA e LSTM tivessem má performance no grupo B.

3.5 Resultados do Grupo C

A série temporal do Grupo C pode ser vista na Figura 36 e os resultados na Tabela 28 e Tabela 29.

O exemplar do modelo C tem um histórico de dados com valores zero nos períodos

Tabela 25 – Grupo A - Desempenho por Modelo

Cluster	Loja	Produto	Modelo	RMSE (hl)	MAPE (%)
A	23	21	Naive	4,16	944,33
A	23	21	Seasonal Naive	1,71	182,69
A	23	21	SARIMA Univariate	-	-
A	23	21	SARIMA Multivariate	-	-
A	23	21	LSTM Univariate	0,91	74,30
A	23	21	LSTM Multivariate	1,13	264,02

Fonte: Elaborada pelo autor.

Tabela 26 – Grupo B - Previsões por Modelo

Mês	Real (hl)	Naive	Seasonal Naive	SARIMA uni	SARIMA multi	LSTM uni	LSTM multi
2017-01-01	2.223	3.848	2.982	3.848	2.469	4.110	4.770
2017-02-01	2.650	3.848	4.175	3.848	4.348	3.854	4.548
2017-03-01	4.182	3.848	4.090	3.848	4.763	3.776	4.527
2017-04-01	4.345	3.848	4.601	3.848	4.770	3.972	4.514
2017-05-01	4.090	3.848	5.027	3.848	4.162	4.128	4.530
2017-06-01	4.004	3.848	4.430	3.848	5.765	4.217	4.519
2017-07-01	4.430	3.848	4.686	3.848	3.863	4.366	4.931
2017-08-01	3.408	3.848	3.664	3.848	3.697	4.529	4.759
2017-09-01	1.534	3.848	2.641	3.848	4.435	4.377	4.647
2017-10-01	4.260	3.848	3.249	3.848	3.500	3.942	4.654
2017-11-01	3.919	3.848	4.019	3.848	3.960	4.049	4.599
2017-12-01	3.152	3.848	3.848	3.848	2.541	4.091	4.526

Fonte: Elaborada pelo autor.

Tabela 27 – Grupo B - Desempenho por Modelo

Cluster	Loja	Produto	Modelo	RMSE (hl)	MAPE (%)
B	26	11	Naive	0,96	29,84
B	26	11	Seasonal Naive	0,76	22,26
B	26	11	SARIMA Univariate	0,96	29,84
B	26	11	SARIMA Multivariate	1,17	32,77
B	26	11	LSTM Univariate	1,14	34,60
B	26	11	LSTM Multivariate	1,44	45,50

Fonte: Elaborada pelo autor.

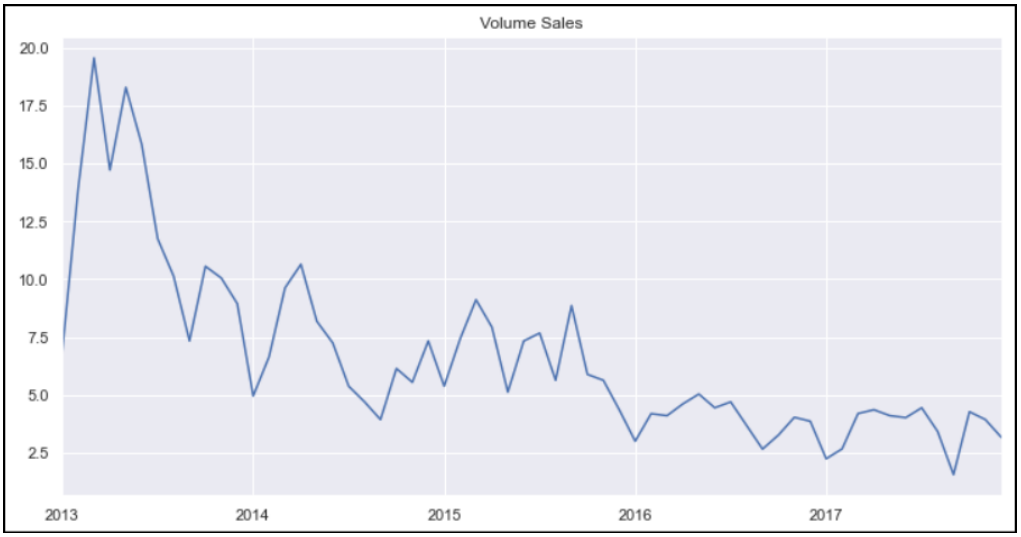

Figura 35 – Série Temporal Grupo B - Loja 26 - Produto 11

Fonte: Elaborada pelo autor

iniciais, e depois assume uma tendência decrescente, sendo que o modelo que melhor capturou esse comportamento foi o SARIMA Multivariado.

Os valores de MAPE não puderam ser calculados pois existe a presença de valor real igual a zero na série.

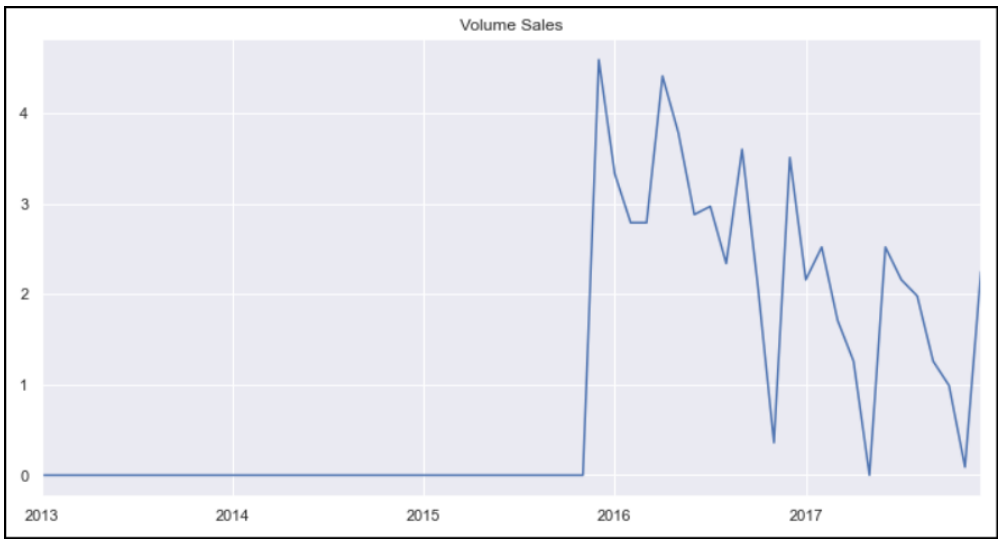

Figura 36 – Série Temporal Grupo C - Loja 60 - Produto 23

Fonte: Elaborada pelo autor

3.6 Resultados do Grupo D

A série temporal do Grupo D pode ser vista na Figura 37 e os resultados na Tabela 30 e Tabela 31.

Tabela 28 – Grupo C - Previsões por Modelo

Mês	Real (hl)	Naive	Seasonal Naive	SARIMA uni	SARIMA multi	LSTM uni	LSTM multi
2017-01-01	2.160	3.510	3.330	2.250	1.638	1.407	3.193
2017-02-01	2.520	3.510	2.790	1.710	1.671	1.774	3.351
2017-03-01	1.710	3.510	2.790	1.710	1.703	1.622	3.081
2017-04-01	1.260	3.510	4.410	3.330	1.740	1.490	2.692
2017-05-01	0.000	3.510	3.780	2.700	-0.138	1.219	2.557
2017-06-01	2.520	3.510	2.880	1.800	3.719	0.812	0.680
2017-07-01	2.160	3.510	2.970	1.890	1.841	1.134	2.730
2017-08-01	1.980	3.510	2.340	1.260	1.875	1.599	2.546
2017-09-01	1.260	3.510	3.600	2.520	1.910	1.493	2.511
2017-10-01	0.990	3.510	2.070	0.990	1.946	1.281	2.345
2017-11-01	0.090	3.510	0.360	-0.720	1.977	1.050	2.609
2017-12-01	2.250	3.510	3.510	2.430	2.004	0.769	2.382

Fonte: Elaborada pelo autor.

Tabela 29 – Grupo C - Desempenho por Modelo

Cluster	Loja	Produto	Modelo	RMSE (hl)	MAPE (%)
C	60	23	Naive	2,11	-
C	60	23	Seasonal Naive	1,73	-
C	60	23	SARIMA Univariate	1,14	-
C	60	23	SARIMA Multivariate	0,80	-
C	60	23	LSTM Univariate	0,91	-
C	60	23	LSTM Multivariate	1,47	-

Fonte: Elaborada pelo autor.

O exemplar do grupo D tem dados durante todo o período analisado e com comportamento e tendências mais bem definidas.

Essa maior quantidade de dados tende a favorecer modelos de maior capacidade de treinamento, e o de melhor performance foi o SARIMA Univariado.

3.7 Resultados do Grupo E

A série temporal do Grupo E pode ser vista na Figura 35 e os resultados na Tabela 32 e Tabela 33.

O exemplar do grupo E possui uma grande quantidade de dados de treinamento disponíveis, favorecendo modelos de maior capacidade de previsão e o de melhor performance foi o SARIMA multivariado.

Tabela 30 – Grupo D - Previsões por Modelo

Mês	Real (hl)	Naive	Seasonal Naive	SARIMA uni	SARIMA multi	LSTM uni	LSTM multi
2017-01-01	2070.900	2843.532	2075.112	2229.103	2884.257	2791.572	2867.912
2017-02-01	2357.640	2843.532	2610.144	2747.569	2725.482	2523.859	3006.387
2017-03-01	2806.812	2843.532	2882.196	3015.401	3032.041	2621.679	2900.695
2017-04-01	3338.388	2843.532	2715.336	2847.466	3005.940	2778.553	3058.551
2017-05-01	3435.156	2843.532	3851.064	3982.921	3615.399	2969.860	3250.798
2017-06-01	3402.108	2843.532	3118.284	3250.071	3652.021	3005.334	3417.611
2017-07-01	2757.132	2843.532	2827.116	2958.885	3258.866	2993.197	3576.837
2017-08-01	3047.544	2843.532	2442.960	2574.725	3201.403	2760.985	3236.240
2017-09-01	3456.432	2843.532	3252.852	3384.615	3365.767	2864.438	3307.710
2017-10-01	3305.556	2843.532	2751.516	2883.279	3132.992	3013.160	3285.088
2017-11-01	2680.884	2843.532	2625.264	2757.027	3339.964	2957.869	3290.017
2017-12-01	2216.592	2843.532	2843.532	2975.295	2481.010	2734.127	3241.983

Fonte: Elaborada pelo autor.

Tabela 31 – Grupo D - Desempenho por Modelo

Cluster	Loja	Produto	Modelo	RMSE (hl)	MAPE (%)
D	32	2	Naive	482,96	15,30
D	32	2	Seasonal Naive	390,10	10,67
D	32	2	SARIMA Univariate	388,79	11,79
D	32	2	SARIMA Multivariate	394,84	12,75
D	32	2	LSTM Univariate	426,91	14,00
D	32	2	LSTM Multivariate	526,78	16,11

Fonte: Elaborada pelo autor.

Tabela 32 – Grupo E - Previsões por Modelo

Mês	Real (hl)	Naive	Seasonal Naive	SARIMA uni	SARIMA multi	LSTM uni	LSTM multi
2017-01-01	11866.4	19214.0	9420.5	14640.4	15910.3	16929.8	17902.3
2017-02-01	15508.9	19214.0	14422.8	17654.0	16573.7	15608.2	18678.7
2017-03-01	18004.1	19214.0	21502.3	21929.7	20917.7	15315.1	18640.0
2017-04-01	16767.2	19214.0	19469.3	20649.1	18931.0	15292.9	18430.7
2017-05-01	21860.0	19214.0	22526.6	22410.3	20667.7	16092.1	18165.0
2017-06-01	19969.7	19214.0	18525.6	20606.6	20971.0	17259.1	19027.2
2017-07-01	21382.1	19214.0	16232.8	19612.0	18137.1	17649.8	19560.8
2017-08-01	20437.6	19214.0	14634.2	19134.5	17020.8	18436.8	19351.9
2017-09-01	20162.3	19214.0	19024.4	20621.3	19139.1	18197.6	18917.1
2017-10-01	17254.8	19214.0	18693.5	20352.9	18482.1	18205.3	18605.0
2017-11-01	17668.8	19214.0	17123.8	18831.2	18403.1	17415.5	18520.5
2017-12-01	18350.9	19214.0	19214.0	21533.1	17542.3	16897.7	18826.0

Fonte: Elaborada pelo autor.

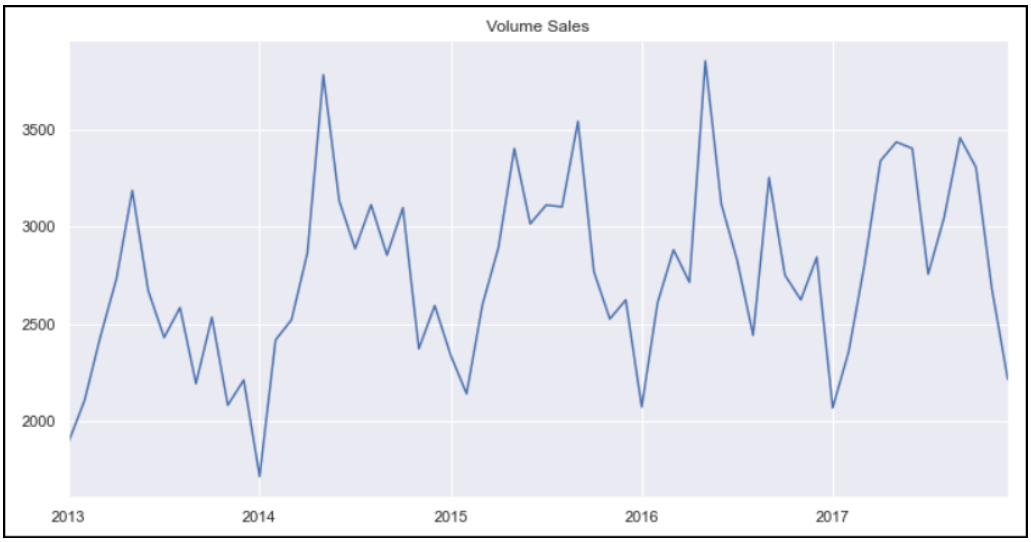

Figura 37 – Série Temporal Grupo D - Loja 32 - Produto 2

Fonte: Elaborada pelo autor

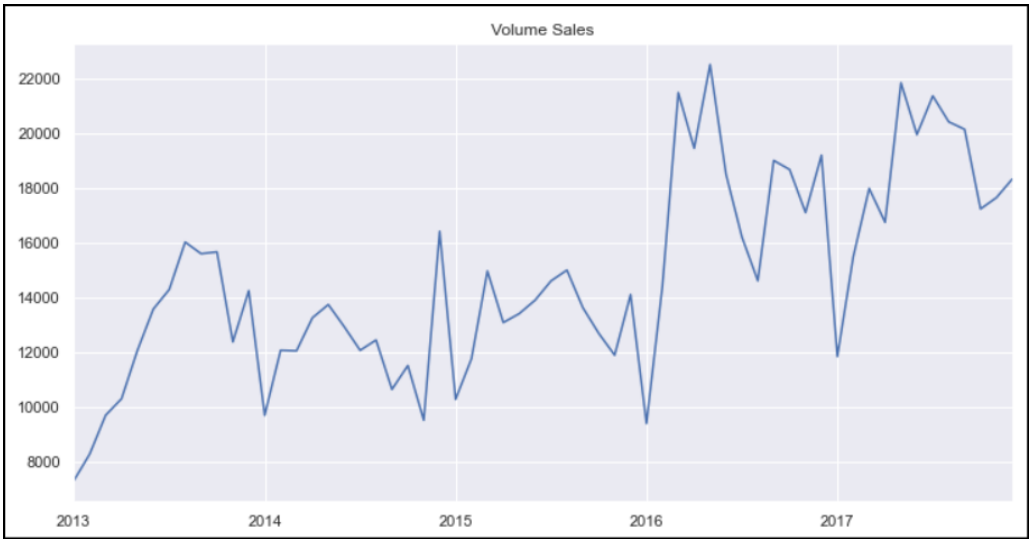

Figura 38 – Série Temporal Grupo E - Loja 2 - Produto 3

Fonte: Elaborada pelo autor

Tabela 33 – Grupo E - Desempenho por Modelo

Cluster	Loja	Produto	Modelo	RMSE (hl)	MAPE (%)
E	2	3	Naive	2840,26	14,05
E	2	3	Seasonal Naive	2796,28	12,31
E	2	3	SARIMA Univariate	2408,35	12,22
E	2	3	SARIMA Multivariate	2217,97	11,10
E	2	3	LSTM Univariate	2894,53	13,24
E	2	3	LSTM Multivariate	2466,66	11,80

Fonte: Elaborada pelo autor.

4 CONCLUSÃO

O resultado encontrado está de acordo com a literatura pesquisada, ou seja, não é possível de antemão definir um modelo que terá o melhor desempenho, sendo necessário modelizar, ajustar e medir para definir qual o modelo mais se adéqua ao objetivo do problema.

No presente estudo, para 5 grupos diferentes de Loja x Produto, foram obtidos 4 modelos diferentes com o melhor desempenho, entre os 6 estudados para a modelagem final.

Inclusive em algumas situações, modelos *Naive* que são simples, rápidos e de alta performance computacional tiveram performance melhor do que modelos mais sofisticados, como o SARIMA e o LSTM.

Possivelmente, a pequena quantidade de dados não possibilitou que os modelos de redes neurais tivessem um melhor desempenho.

Outro aspecto a se analisar é a quantidade de variáveis preditoras, sendo que em apenas dois grupos o modelo Multivariado teve melhor desempenho.

Uma desvantagem do modelo Multivariado nessa caso é que as variáveis preditoras não são conhecidas no futuro, como por exemplo temperatura, população e renda, e teriam que ser usadas estimativas em lugar de valores reais, aumentando a incerteza sobre a previsão da variável objetivo.

O desempenho medido por MAPE não pode ser medido em um grupo devido à presença de valores reais iguais a zero, e teve uma tendência de valores maiores para magnitudes mais baixas da variável predita com valores próximos a zero, demonstrando a limitação do seu uso e que está perfeitamente alinhada à literatura.

Os indicadores de desempenho TU e POCID, usados na fase de seleção dos modelos e que são específicos para séries temporais, mostraram o seu valor na comparação de modelos em complemento aos indicadores tradicionais RMSE e MAPE.

Como próximos passos para melhorar o desempenho dos modelos, pode-se explorar uma ou mais das técnicas abaixo:

- Fazer *oversampling* e reaplicar o ajuste fino dos modelos, especialmente as redes neurais LSTM que necessitam de uma grande quantidade de dados para seu treinamento;
- Explorar a otimização de mais hiper-parâmetros do SARIMA e LSTM, incluindo o número de neurônios e camadas do LSTM;

- Incluir os dados do calendário de eventos e feriados como variáveis exógenas;
- Utilizar a primeira diferença das variáveis para a modelização LSTM, a exemplo do que foi feito no SARIMA;
- Testar técnicas de regressão que consideram os registros como independentes para a tarefa de predição fora da faixa de treinamento.

REFERÊNCIAS

ARMSTRONG, J. S. **Principles of forecasting: a handbook for researchers and practitioners**. [S.l.]: Springer, 2001. v. 30.

CHEN, I.-F.; LU, C.-J. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. **Neural Computing and Applications**, Springer, v. 28, n. 9, p. 2633–2647, 2017.

CHERIYAN, S. et al. Intelligent sales prediction using machine learning techniques. In: IEEE. **2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)**. [S.l.], 2018. p. 53–58.

COOPER, L. G. et al. Promocast™: A new forecasting method for promotion planning. **Marketing Science**, INFORMS, v. 18, n. 3, p. 301–316, 1999.

CORSTEN, D.; GRUEN, T. Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. **International Journal of Retail & Distribution Management**, MCB UP Ltd, 2003.

GALLAGHER, C.; MADDEN, M. G.; D'ARCY, B. A bayesian classification approach to improving performance for a real-world sales forecasting application. In: IEEE. **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. [S.l.], 2015. p. 475–480.

GOOIJER, J. G. D.; HYNDMAN, R. J. 25 years of time series forecasting. **International journal of forecasting**, Elsevier, v. 22, n. 3, p. 443–473, 2006.

HERBIG, P. A.; MILEWICZ, J.; GOLDEN, J. E. The do's and don'ts of sales forecasting. **Industrial Marketing Management**, Elsevier, v. 22, n. 1, p. 49–57, 1993.

HUANG, T.; FILDES, R.; SOOPRAMANIEN, D. Forecasting retailer product sales in the presence of structural change. **European Journal of Operational Research**, Elsevier, v. 279, n. 2, p. 459–470, 2019.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. OTexts, 2018. Disponível em: <https://otexts.com/fpp2/index.html>. Acesso em: 29 junho 2021.

JIANG, L. et al. Demand forecasting for alcoholic beverage distribution. **SMU Data Science Review**, v. 3, n. 1, p. 5, 2020.

KRISHNA, A. et al. Sales-forecasting of retail stores using machine learning techniques. In: IEEE. **2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)**. [S.l.], 2018. p. 160–166.

KUO, R.; TSENG, Y.; CHEN, Z.-Y. Integration of fuzzy neural network and artificial immune system-based back-propagation neural network for sales forecasting using qualitative and quantitative data. **Journal of Intelligent Manufacturing**, Springer, v. 27, n. 6, p. 1191–1207, 2016.

LOUREIRO, A. L.; MIGUÉIS, V. L.; SILVA, L. F. da. Exploring the use of deep neural networks for sales forecasting in fashion retail. **Decision Support Systems**, Elsevier, v. 114, p. 81–93, 2018.

MENTZER, J. T.; MOON, M. A. **Sales forecasting management: a demand management approach**. [S.l.]: Sage Publications, 2004. 9 p.

_____. **Sales forecasting management: a demand management approach**. [S.l.]: Sage Publications, 2004. 8 p.

MERKURYEVA, G.; VALBERGA, A.; SMIRNOV, A. Demand forecasting in pharmaceutical supply chains: A case study. **Procedia Computer Science**, Elsevier, v. 149, p. 3–10, 2019.

PAVLYSHENKO, B. M. Machine-learning models for sales time series forecasting. **Data**, Multidisciplinary Digital Publishing Institute, v. 4, n. 1, p. 15, 2019.

SAGAERT, Y. R. et al. Tactical sales forecasting using a very large set of macroeconomic indicators. **European Journal of Operational Research**, Elsevier, v. 264, n. 2, p. 558–569, 2018.

SINGH, D. et al. Machine learning based business forecasting. **IJ Information Engineering and Electronic Business**, v. 6, p. 40–51, 2018.