UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CAIO HENRIQUE RIBEIRO GARCIA DE MEDEIROS

ANÁLISE DO COMPORTAMENTO DO INDICADOR DE VOLUME DE VENDAS COM VARIAÇÃO NO TEMPO

CAMPINA GRANDE - PB

2022

CAIO HENRIQUE RIBEIRO GARCIA DE MEDEIROS

ANÁLISE DO COMPORTAMENTO DO INDICADOR DE VOLUME DE VENDAS COM VARIAÇÃO NO TEMPO

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Melina Mongiovi Cunha Lima Sabino

CAMPINA GRANDE - PB

2022

CAIO HENRIQUE RIBEIRO GARCIA DE MEDEIROS

ANÁLISE DO COMPORTAMENTO DO INDICADOR DE VOLUME DE VENDAS COM VARIAÇÃO NO TEMPO

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Melina Mongiovi Cunha Lima Sabino

Orientador – UASC/CEEI/UFCG

Eanes Torres Pereira

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de Setembro de 2022.

CAMPINA GRANDE - PB

RESUMO

No mundo dos negócios, a análise de indicadores para tomada de decisões é muito importante . Um indicador comumente avaliado é o de *Volume de Vendas*, o qual pode ser decomposto por mercado, linha de produto, meios de distribuição etc. Com base nesse indicador, pode-se compreender como as vendas da empresa se comportam e, então, decidir que ações deverão ser tomadas para aumentar os lucros ou reduzir prejuízos de uma empresa. Desta forma, este estudo teve como objetivo analisar o comportamento do indicador *Volume de Vendas* de algumas empresas vendedoras de café, para identificar possíveis correlações e tendências entre mercados e estimar o *Volume de Vendas* para o ano seguinte. Para isto, foram aplicadas algumas técnicas de extração e análise de dados, de correlação e de auto regressão, para predição deste indicador de vendas. Na realização do trabalho, foi utilizado uma base de dados de quatro empresas vendedoras de café, clientes de uma empresa de consultoria em análise de dados.

Análise do comportamento do indicador de Volume de Vendas com variação no tempo

Caio Henrique R. G. de Medeiros
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
caio.medeiros@ccc.ufcg.edu.br

Profª.Drª.Melina Mongiovi C. L. Sabino
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
melina@computacao.ufcg.edu.br

RESUMO

No mundo dos negócios, a análise de indicadores para tomada de decisões é muito importante . Um indicador comumente avaliado é o de *Volume de Vendas*, o qual pode ser decomposto por mercado, linha de produto, meios de distribuição etc. Com base nesse indicador, pode-se compreender como as vendas da empresa se comportam e, então, decidir que ações deverão ser tomadas para aumentar os lucros ou reduzir prejuízos de uma empresa. Desta forma, este estudo teve como objetivo analisar o comportamento do indicador *Volume de Vendas* de algumas empresas vendedoras de café, para identificar possíveis correlações e tendências entre mercados e estimar o *Volume de Vendas* para o ano seguinte. Para isto, foram aplicadas algumas técnicas de extração e análise de dados, de correlação e de auto regressão, para predição deste indicador de vendas. Na realização do trabalho, foi utilizado uma base de dados de quatro empresas vendedoras de café, clientes de uma empresa de consultoria em análise de dados.

Palavras-chave

técnica, análise, dados, comportamento, correlação, autoregressão

ABSTRACT

In the business world, it is very important to analyze indicators for decision making. A commonly analyzed indicator is Sales Volume, which can be broken down by market, product line, means of distribution, etc. Based on this indicator, one can have clarity on how the company's sales behave and, then, decide what actions should be taken to increase a company's profits or reduce losses. In this way, this study aimed to analyze the behavior of the "Sales Volume" indicator of some coffee selling companies, to identify possible correlations and trends between markets and to estimate the sales volume for the following year. Some techniques of data extraction, data analysis, correlation and auto-regression were applied to analyze and predict the sales indicator. Carrying out the work, a database of four coffee sales companies, which were clients of a data analysis consulting company, was used.

Keywords

technique, analysis, data, behavior, correlation, autoregression.

1. INTRODUÇÃO

O café é um dos produtos mais consumidos no mundo. Segundo dados da Organização Internacional de Café (do inglês, *International Coffee Organization* - ICO) [9], houve um consumo de 167,58 milhões de sacas de 60 Kg de café em 2020, representando um aumento de 1,9% em relação ao ciclo anterior. A Tabela [1] apresenta os valores do consumo, em milhares de sacas de 60Kg, no mundo e por regiões.

	2017 -2018	2018-2019	2019-2020	2020-2021
Mundo	**161.377**	**168.492**	**164.202**	**166.346**
África	11.087	12.017	12.024	12.242
Ásia e Oceania	34.903	36.472	36.002	26.503
América Central e México	5.273	5.432	5.327	5.364
Europa	53.251	55.637	53.372	54.065
América do Norte	29.941	31.779	30.580	30.993
América do Sul	26.922	27.156	26.898	27.180

Tabela [1] - Consumo Mundial de Café, em milhares de sacas de 60 Kg. (Fonte: *International Coffee Organization*, 2021)

Diversos países mantêm altas taxas de importação deste produto. Ainda de acordo com os dados da ICO, a Europa - considerando os países da União Européia e alguns outros que não pertencem ao conselho - importou mais de 88 milhões de sacas de 60Kg, entre dezembro de 2020 e novembro de 2021. Dentro deste mercado, os maiores consumidores são Alemanha, Itália, França e Bélgica.

Assim como existem os grandes importadores e consumidores de café, há também os grandes produtores. O Brasil

é o maior produtor e exportador de café do mundo, seguido do Vietnã e da Colômbia. A Tabela [2] apresenta o ranking das maiores exportações entre os anos de 2020 a 2021 e 2021 a 2022.

	2020-2021	2021-2022
1º Brasil	39.076, 368	44.847,548
2º Vietnã	25.932,706	25.625,000
3º Colômbia	12.495,209	12.498,995
4º Indonésia	7.106,838	7.277,934
5º Uganda	6.725,408	5.465,723

Tabela [2] - Ranking dos 5 maiores países exportadores de café, em milhares de sacas de 60Kg.
(Fonte: *International Coffee Organization*, 2022)

As companhias que atuam no ramo de vendas de café precisam, como qualquer empresa, ter controle sobre suas vendas para poderem melhorar seus ganhos. Indicadores chaves de desempenho (do inglês *Key Performance Indicators - KPI*), ou apenas indicadores de desempenho, "são ferramentas modernas que ajudam a manter o desempenho do setor produtivo em alto nível" [11]. De acordo com a Fundação Nacional da Qualidade (FNQ) [5], a existência de um bom sistema de indicadores de desempenho em uma organização permite uma análise muito mais profunda e abrangente sobre a efetividade da gestão e de seus resultados. Para que as empresas vendedoras de café possam ter uma melhor gestão sobre suas vendas por meio dos indicadores, elas precisam obter os dados que irão compô-los.

Algumas das dificuldades comuns para as empresas em relação aos seus indicadores encontram-se na coleta e manipulação dos dados para a extração das informações desejadas. A qualidade dos dados ainda é uma das maiores preocupações para as empresas e dados "sujos" levam a decisões incorretas e análises não confiáveis. Exemplos de erros comuns incluem valores não existentes, erros de digitação, formatos de dados misturados, entradas de valores repetidas de uma mesma entidade e violação das regras de negócio [2]. Para evitar tais problemas, é importante realizar a limpeza dos dados (do inglês, *Data Cleansing* ou *Data Cleaning*), que consiste em uma operação realizada sobre os dados, no intuito de remover anomalias e obter um conjunto de dados que seja uma representação única e precisa do mundo real, numa versão simplificada [15]. Este processo envolve remover erros, resolver inconsistências e transformar os dados em um formato uniforme [7]. Uma vez tratados, pode-se iniciar a análise dos dados.

O presente trabalho teve como objetivo limpar, reestruturar e filtrar uma base de dados de quatro empresas vendedoras de café, clientes de uma empresa de consultoria em análise de dados, e realizar análises de um de seus indicadores de vendas, o *Volume de Vendas*, com o intuito de identificar possíveis correlações e tendências entre os diferentes mercados que estas empresas atendem. Além disso, no intuito de tentar estimar o *Volume de Vendas* de cada mercado para o ano seguinte, foram testados alguns modelos de Autoregressão.

A seção 2 deste documento expõe a Fundamentação Teórica de alguns conceitos pertinentes à pesquisa. A seção 3 discorre sobre a Metodologia aplicada no estudo. A seção 4 apresenta os Resultados deste trabalho e a seção 5, a Conclusão.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Indicador de Desempenho e Volume de Vendas

Um indicador de desempenho é uma informação quantitativa ou qualitativa que expressa o desempenho de um processo, em termos de eficiência, eficácia ou nível de satisfação e que, em geral, permite acompanhar sua evolução ao longo do tempo e compará-lo com outras organizações [5]. Tem sido mostrado, com o passar dos anos, que um KPI pode ser composto por diversos outros KPIs. O KPI de satisfação de cliente, por exemplo, pode ser composto por tempo, custo, qualidade e comunicação eficiente na relação com o cliente [16]. Desta forma, pode-se definir dois tipos de KPIs: o atômico e o composto. KPIs atômicos são aqueles obtidos diretamente dos dados, como "Número de veículos vendidos". Já os KPIs compostos são formados por outros KPIs [12]. O *Volume de Vendas* é um KPI atômico, podendo também compor outros indicadores.

2.2 Séries Temporais

O *Volume de Vendas*, objeto de estudo deste trabalho, apresenta alterações em seu valor ao longo do tempo. Esta é uma das características de uma série temporal. Uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo [4]. De modo geral, os valores próximos são dependentes e o estudo de séries temporais objetiva analisar e modelar estas dependências. Portanto, a ordem dos dados é fundamental para as análises neste tipo de série.

Características comuns a serem analisadas nas séries temporais são tendência, ciclo e sazonalidade, de acordo com Morettin & Toloi [14]. Contudo, para definição de qual modelo é mais adequado e quais parâmetros podem trazer melhores resultados, é importante também realizar a análise de estacionariedade de uma série temporal. De acordo com Wooldridge [31] um processo estacionário é aquele que possui a distribuição de probabilidades estáveis ao longo do tempo. Desta forma, pode-se determinar que uma série temporal é estacionária quando suas propriedades estatísticas, como média, variância e autocorrelação, mantém-se constantes com o passar do tempo. Esse estudo realizou testes de sazonalidade e estacionariedade para aplicação dos modelos de auto regressão. Com a produção de alguns gráficos, também foi possível perceber certas tendências de crescimento ou declínio em alguns mercados.

2.3 Modelos de Autoregressão

O modelo de autoregressão é um modelo estatístico utilizado para prever valores futuros de uma série temporal. Ele utiliza os valores históricos da série e estima valores no futuro, baseando-se na dependência existente entre os valores passados. Existem diversos modelos e técnicas de autoregressão. Neste trabalho foram utilizadas algumas técnicas de autoregressão, apresentadas a seguir.

2.3.1 Autoregressão Simples

Modelos Autoregressivos são uma classe de modelos probabilísticos que modelam a distribuição de dados ao estimar a densidade destes dados [3] . Eles são muito populares em algumas áreas, como em Economia, onde é natural pensar o valor de alguma variável no instante t como função de valores defasados da mesma variável [14]. Este modelo de auto regressão se baseia nos valores passados para estimar o seguinte, seguindo a Equação [1] , de acordo com Gujarati & Porter [8]:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_k X_{t-k} + u_t$$

Equação [1] - Cálculo do modelo autoregressivo de defasagem distribuída.

Nesta equação, o alfa é uma constante, o coeficiente beta zero é conhecido como multiplicador de curto prazo, porque dá a variação do valor médio de Y em decorrência da variação unitária de X no mesmo período e u é o termo de erro. Depois de k períodos, o somatório dos betas é chamado de multiplicador de defasagens de longo prazo [8].

2.3.2 AR, MA, ARMA e ARIMA

Os modelos Autoregressivos (AR) e Média Móvel (MA) são baseados na suposição de que a série temporal seja gerada através de um sistema linear, e que possuem um termo de erro aleatório não correlacionado, com média zero e variância constante, ou seja, um ruído branco [19]. É possível que um processo tenha tanto características AR quanto MA e seja, portanto, ARMA (*Autoregressive Moving Average*). Desta forma, o modelo deve apresentar termos autoregressivos p e termos de média móvel q. [8]

O método ARIMA (*Autoregressive Integrated Moving Average*) é um dos mais populares modelos de análise de previsão para séries temporais [19]. Quando as séries temporais não são estacionárias, elas são consideradas integradas. Se tivermos de diferenciar uma série temporal d vezes para torná-la estacionária e aplicar-lhe o modelo ARMA (p, q), diremos que a série temporal original é ARIMA (p, d, q), ou seja, ela é uma série temporal autoregressiva integrada de médias móveis, em que p denota os números dos termos autoregressivos, d o número de vezes que a série deve ser diferenciada antes de tornar-se estacionária e q o número de termos de média móvel [8].

3. METODOLOGIA

3.1 Caracterização da base de dados

Neste estudo foi utilizada uma base de dados constituída por dados de quatro empresas vendedoras de café, clientes de uma consultoria especializada em análise de dados e *Big Data* que os forneceu. Por questões de confidencialidade, nem todas as informações presentes no conjunto de dados puderam ser detalhadas neste trabalho. Esta base de dados foi disponibilizada em formato ".csv". Os atributos principais estudados da base de dados original foram **Empresa**, **Canal de Distribuição**, **Mercado** e **Valor de Ocorrência**, que é *Volume de Vendas*, dado em toneladas (t), com ocorrências mensais. Os dados presentes ocorreram de 01 de janeiro de 2012 a 01 de janeiro de 2022.. O indicador era registrado no primeiro dia de cada mês e o conjunto de dados inicial possuía 8938 ocorrências.

3.2 Execução do estudo e observações sobre o conjunto de dados

O estudo foi iniciado com uma manipulação e preparação dos dados. Dados inconsistentes e faltantes (*Not a Number* - NaN) foram removidos. Em seguida, o conjunto dos dados foi filtrado, para conter apenas as quatro informações que foram foco deste estudo.

Realizou-se uma análise exploratória sobre o conjunto de dados filtrados para se obter um melhor entendimento sobre eles, através da aplicação de agrupamentos nos atributos que eram foco deste estudo. Os agrupamentos foram realizados sobre o *Dataframe* gerado com a base de dados original, utilizando-se a biblioteca Pandas [20]. Os dados foram filtrados inicialmente por Empresa, Canal de Distribuição e Mercado, Valor de Ocorrência (*Volume de Vendas*) e a Data. O Atributo Valor de Ocorrência foi renomeado *Volume de Vendas*, as siglas referentes às empresas foram substituídas por E1, E2, E3 e E4 e as dos canais por C1, C2 e C3. Mantiveram-se os nomes originais dos mercados, sendo um deles, chamado de 'Indefinido', desconsiderado das análises.

Os seis mercados estudados foram: América do Norte (AN), Ásia (AS), Europa (EU), América Latina (AL), Japão (JP) e Brasil (BRL), mas não foi fornecida nenhuma informação adicional sobre os países que compunham alguns destes mercados.

Percebeu-se que as empresas não atuavam em todos os canais de distribuição e nem em todos os mercados. A Tabela [3] mostra a distribuição de canais e mercados para cada empresa. Desta forma, após a filtragem inicial dos dados, quatro notebooks do Jupyter [10] foram criados separadamente, para a preparação individual de cada empresa. O objetivo era ordenar os dados por data, de cada canal e cada mercado da empresa em questão, juntá-los em um novo *Dataframe* e gerar um arquivo '.csv' com os *Volumes de Vendas* individuais de cada empresa.

Para a análise dos mercados, um novo notebook foi gerado. Nele, foram realizadas as análises de *Volume de Vendas* médio de cada 'Empresa', de cada 'Canal', de cada 'Mercado' e de cada 'Empresa, por Canal e Mercado'. Por fim, gerou-se um novo *Dataframe* com o somatório total de cada um dos seis mercados em estudo, para o início das análises das séries temporais e da autoregressão.

Percebeu-se que os valores de cada mercado não ocorriam todos no mesmo período de tempo. Os mercados AN, AS, EU, JP e BRL, quando somados com todas suas ocorrências, apresentavam valores de 2012 a 1 de janeiro de 2022. Esta ocorrência isolada de 2022 foi desconsiderada nas análises. Contudo, o mercado AL apresentava ocorrência de *Volume de Vendas* apenas até 01 de janeiro de 2016. Portanto, para realizar a predição com auto regressão para os mercados AN, AS, EU, JP e BRL, foram utilizados os valores de 2012 a 2020 para treino dos modelos, teste em 2021 e predição de 2022. Para o mercado AL, utilizaram-se os valores de 2012 a 2014 para treino, teste em 2015 e predição de 2016. Esta quantidade reduzida de dados afetou os resultados, como será visto mais a frente. Em seguida, foram realizadas as análises de correlação e de regressão nos mercados.

Antes da aplicação das quatro técnicas de autoregressão, foi feito também uma breve análise de autocorrelação, com o módulo pacf da biblioteca statsmodel [29], e testes de estacionariedade, utilizando-se a biblioteca pmdarima [23] e um de seus módulos chamado autoARIMA, que realiza testes com diferentes parâmetros sobre a série temporal e retorna a sugestão dos melhores valores para p, d, e q.

Empresa	Canal de Distribuição	Mercado
E1	C1	AN
E2	C3	AN
		AS
		EU
		AL
		JP
	C1	Indefinido
	C2	BRL
E3	C1	AS
		EU
		JP
E4	C1	AS
		AL
		JP

Tabela [3] - Canais e Mercados de atuação de cada empresa.

A aplicação dos modelos de autoregressão foi iniciada pela Autoregressão Simples. Este modelo possui um parâmetro chamado *lag*, que define qual será o intervalo de tempo onde os resultados de um período de tempo afetará os seguintes. O valor foi testado no intervalo de [1, 6] para se obter o melhor resultado e os dados de todos os seis mercados foram testados neste modelo.

Quanto ao modelo ARIMA e suas variantes (ARMA, AR e MA), houveram dois momentos de testes. No primeiro momento aplicou-se o modelo ARIMA sobre cinco mercados, utilizando-se os parâmetros (p, d, q) recomendados pelo módulo autoARIMA. Como o mercado AL obteve os parâmetros (0,0,0) como recomendação, ele foi desconsiderado desta análise.

A Tabela [4] apresenta os parâmetros recomendados para cada mercado pelo autoARIMA. É importante mencionar que para este modelo ARIMA, dependendo de como se varia os parâmetros *p*, *d* e *q*, obtém-se modelos diferentes. Para executar um modelo AR, por exemplo, o modelo deve ser executado apenas com o parâmetro *p*, como em (1, 0, 0). Para um modelo MA, executa-se o modelo apenas com o parâmetro *q* ativo (0, 0, 1). Para a execução do modelo ARIMA, é preciso que os 3 parâmetros sejam diferentes de 0.

Após a execução com os parâmetros recomendados pelo autoARIMA, percebeu-se que alguns mercados não tiveram bons resultados, principalmente para os casos em que o modelo ARIMA era aplicado. Decidiu-se, portanto, realizar um novo teste, com uma combinação de diversos valores para *p*, *d* e *q*, numa espécie de tentativa e erro, para se obter os melhores valores possíveis do r^2 e MAE para cada mercado. Por fim, foi testado o módulo Regression do modelo XGBoost - uma biblioteca que aplica uma versão melhorada do *Gradient Boost* e tem um módulo para regressão [30] - e registraram-se as métricas obtidas para cada um dos seis mercados. Neste último modelo o mercado AL voltou a ser considerado.

Mercado	(p,d,q) recomendado	Modelo a executar
AN	(0,1,1)	(IMA)
AS	(0,1,2)	(IMA)
EU	(0,1,1)	(IMA)
AL	(0,0,0)	-
JP	(2,0,0)	AR
BRL	(0,1,1)	IMA

Tabela [4] - Parâmetros recomendados pelo autoARIMA.

3.3 Avaliação dos modelos de autoregressão

Para avaliação das técnicas de auto regressão, foram utilizadas duas medidas básicas: o r^2 e o erro médio absoluto (do inglês, *Mean Absolute Error* - MAE). O coeficiente de determinação, também chamado de r^2, é interpretado como sendo a proporção da variabilidade dos Y's observados, explicada pelo modelo considerado. Seu valor pertence ao intervalo [0,1], sendo que quanto mais próximo a 1, melhor o ajuste do modelo em questão [1].

O Erro Médio Absoluto é uma das formas de medir a distância entre dois vetores: o vetor das previsões e o vetor dos valores alvos [6]. É uma medida de desempenho muito utilizada por não ser muito influenciada por *outliers* (valores muito fora do padrão). Os dados do estudo apresentam alguns outliers em alguns mercados, e portanto, optou-se por utilizar esta métrica. Quanto menor o valor desta métrica, melhor foi o desempenho do modelo.

Ao final de todo o processo, foram geradas predições para os volumes de vendas do ano de 2016 para o mercado AL e para o ano de 2022 para os demais mercados, utilizando-se o modelo Prophet [24], do Facebook, e plotados todos os gráficos.

3.4 Ferramentas utilizadas nas análises

Para preparação do conjunto de dados, análise dos dados e realização da auto regressão foram utilizadas a linguagem de programação Python3 - na versão 3.10 - e as bibliotecas, Pandas [20], *numpy* [18], *Scikit-learn* [26], além das bibliotecas para plotagem e edição de gráficos, matplotplib [13] e *seaborn* [27]. Como já mencionado, também utilizou-se o módulo AutoARIMA, da biblioteca pmdarima [23], para testar diversas opções de parâmetros para o modelo ARIMA e selecionar aquele que obteve melhor resultado. Para o teste de estacionariedade das séries temporais foi utilizado o pacote adfuller, da biblioteca *statsmodel* [29]. Por fim, além dos modelos de auto regressão do *Scikit-learn* e do *statsmodel*, também utilizou-se a biblioteca *Prophet* [24], do Facebook, para predição dos volumes de vendas de 2022 e 2016, para seus respectivos mercados.

O ambiente de desenvolvimento utilizado foi o Jupyter Notebook [10], integrado com o programa VS Code, no sistema operacional Windows 11.

4. RESULTADOS E DISCUSSÃO

4.1 Visão geral do volume de vendas

A empresa que mais vendeu café foi a E2. Contudo, esta é a empresa que possui presença em mais canais e mais mercados,

justificando seu alto *Volume de Vendas*. O Gráfico [1] detalha o volume médio de vendas de cada empresa.

Dos canais de distribuição, o que apresentou o maior índice de *Volume de Vendas* foi o Canal C1. O Gráfico [2] apresenta o volume médio de vendas de cada canal. Com relação aos mercados, o mercado com maior *Volume de Vendas* foi o Europeu, seguido pela Ásia e pelo Japão. O Gráfico [3] apresenta o *Volume de Vendas* médio de todos os mercados. Estes dados mostraram-se de acordo com as informações anteriores, onde foi apresentado que a Europa é o maior mercado consumidor de café da atualidade.

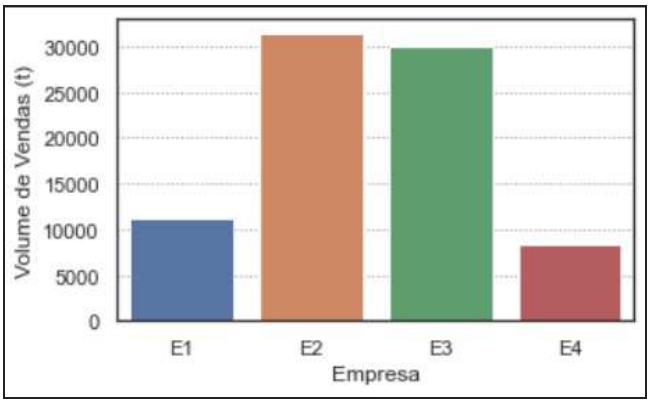

Gráfico [1] - Volume médio de venda de café, por empresa, em toneladas (t).

Para as Empresas E2, E3 e E4. foi feito uma análise do *Volume de Vendas* para seus canais que possuem mais de um mercado. Pelo Gráfico [4], percebe-se que o mercado da Europa apresentou o maior *Volume de Vendas* no Canal C2 da Empresa E2, bem como no Canal C1 da Empresa E3, como visto no Gráfico [5]. Já no Gráfico [6], observa-se que o maior *Volume de Vendas* da Empresa E3, no Canal C1, foi o Japão, um mercado que vem crescendo no consumo do café.

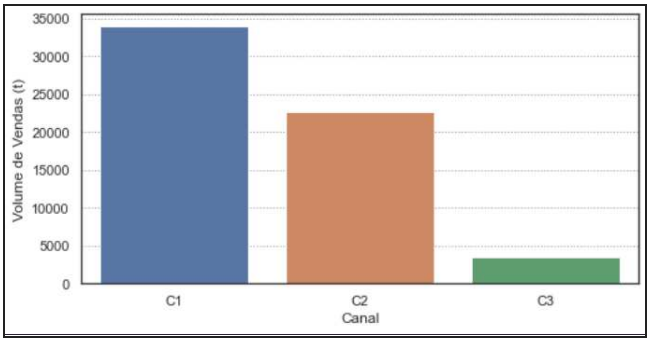

Gráfico [2] - Volume médio de venda de café, por canal, em toneladas (t).

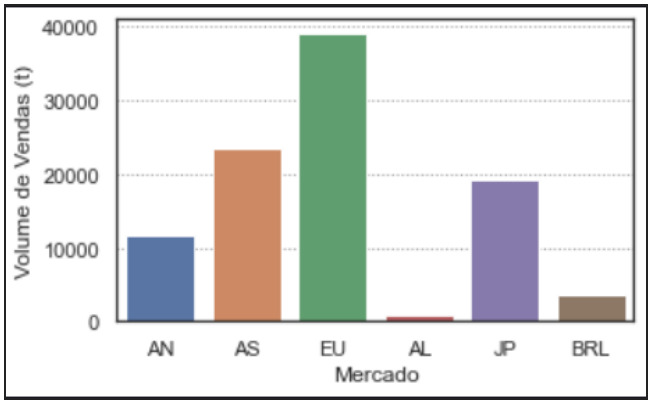

Gráfico [3]- Volume médio de venda de café, por mercado, em toneladas (t).

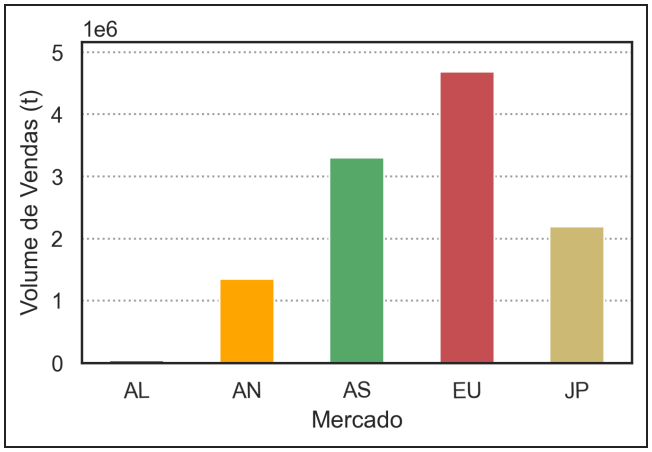

Gráfico [4] - Volume total de venda de café, da empresa E2, no canal 2, por mercado, em toneladas (t).

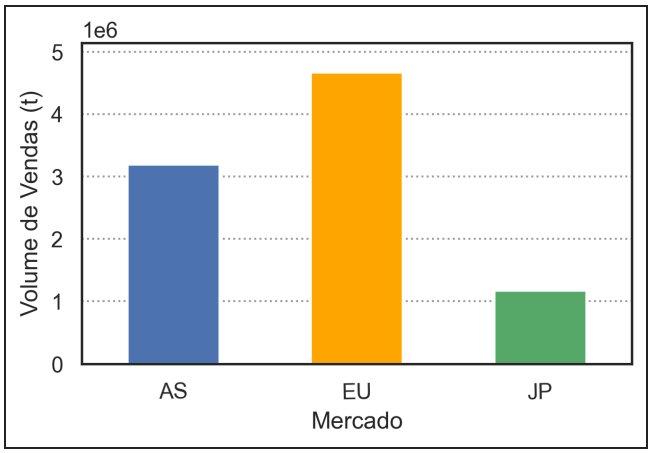

Gráfico [5] - Volume total de venda de café, da empresa E3, no canal 1, por mercado, em toneladas (t).

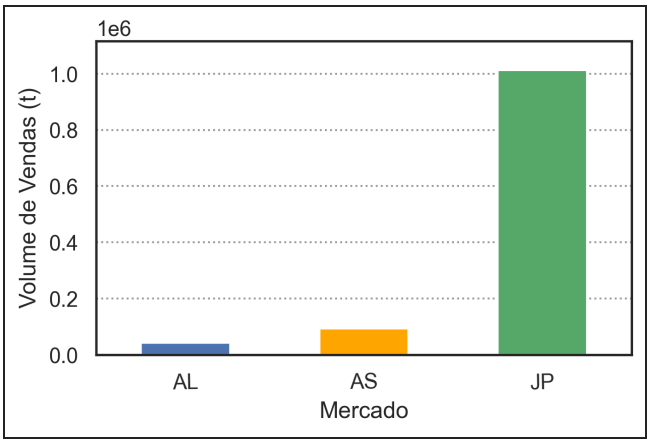

Gráfico [6] - Volume total de venda de café, da empresa E4, no canal 1, por mercado, em toneladas (t).

4.2 Análise de correlação entre mercados

Não foi possível identificar correlações significativas entre os mercados, como apresentado no Gráfico [7]. Apenas dois casos de correlação entre os mercados chamaram a atenção. Percebeu-se uma correlação positiva forte entre os mercados do Brasil (BRL) e da América Latina (AL), como era de se esperar, pois são mercados que estão entre os grandes produtores de café, tendo grande exportação, porém baixa compra interna. Porém, é importante destacar que estes são dois mercados com estacionariedade diferentes, mas como o mercado AL possui dados apenas até 2016, a correlação encontrada refere-se apenas a este período nos dois mercados, onde a tendência de declínio do mercado BRL ainda não estava tão evidente.

Outra correlação que chamou atenção, apesar de ser fraca, foi entre o mercado Europeu (EU) e o Brasileiro (BRL). Mesmo sendo mercados diferentes, com climas opostos e tendências de compras diferentes, houve uma correlação positiva.

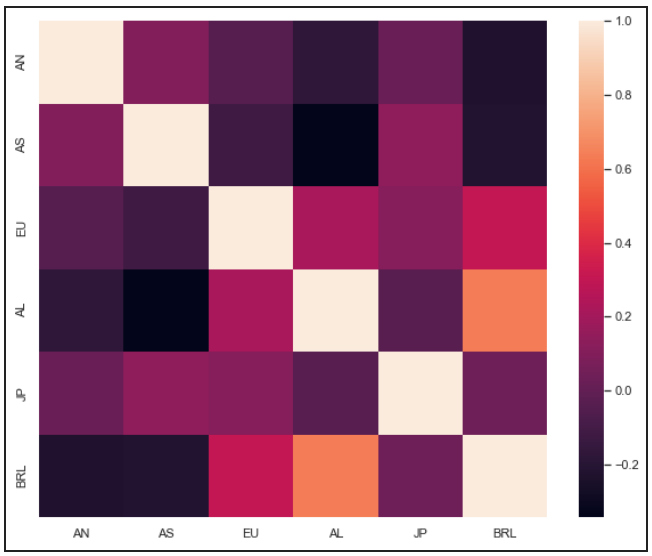

Gráfico [7] - Matriz de correlação entre os mercados.

4.3 Análise das séries temporais

Analisando-se o Gráfico [8], percebe-se que o mercado com maiores volumes de vendas é o da Europa. Entretanto, o mercado da Ásia vem crescendo bastante nos últimos anos. Outro mercado que apresentou uma tendência de crescimento foi o mercado da América do Norte. O mercado do Japão mostrou-se estacionário, mas com volumes de vendas superiores ao do mercado da América do Norte. Por fim, como já foi discutido previamente, os dois menores mercados são os do Brasil e o da América Latina. O mercado do Brasil mostrou um leve declínio a partir de 2017 e o mercado da América Latina não possui dados de *Volume de Vendas* após 2016, como também já foi mencionado.

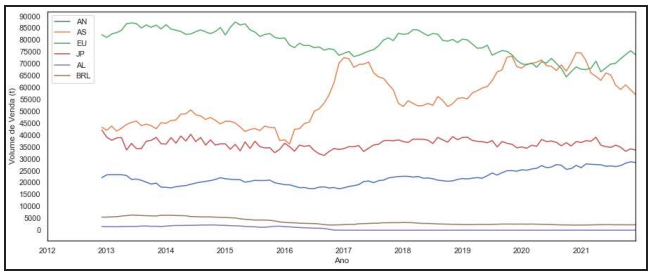

Gráfico [8] - Série temporal das médias móveis do volume de vendas de cada mercado.

4.3.1 Sazonalidade

De modo geral, o estudo da sazonalidade não mostrou padrões sazonais nas séries temporais do *Volume de Vendas* dos mercados, sugerindo uma aleatoriedade significativa na ocorrência dos seus valores. Os Gráfico [9] e Gráfico [10] ilustram a ausência de sazonalidade nos mercados EU e JP. Como a sazonalidade foi calculada pela diferença sucessiva de valores, valores negativos indicam uma queda entre o *Volume de Vendas* do mês anterior e o do mês seguinte.

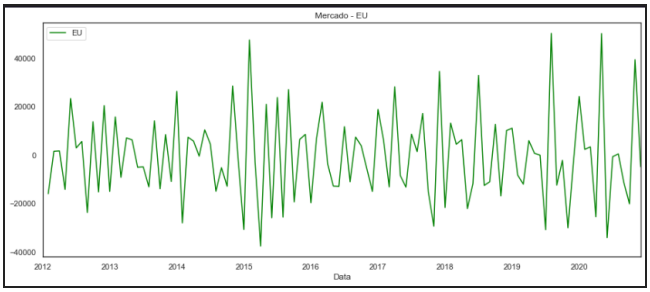

Gráfico [9] - Sazonalidade do mercado EU.

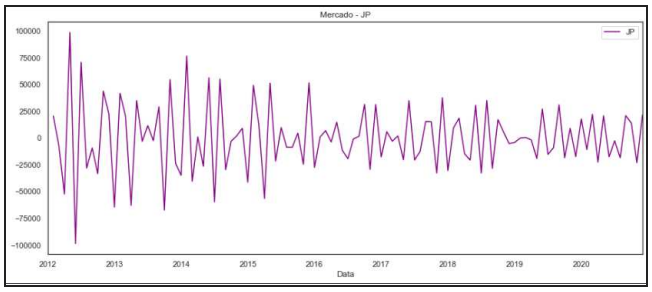

Gráfico [10] - Sazonalidade do mercado JP.

4.3.2 Estacionariedade

O estudo de estacionariedade mostrou que os mercados EU, JP e AL são **estacionários** e os mercados AN, AS e BRL são **não-estacionários**, ou seja, possuem um certa tendência em seu comportamento. Isto corrobora as análises anteriores, onde foi confirmado que os mercados AN e AS mostraram uma tendência de crescimento nos últimos anos e o mercado BRL mostrou uma leve tendência de declínio em relação ao *Volume de Vendas*, a partir de 2017.

4.4 Auto regressão e predição de valores

Segundo o estudo de estacionariedade, como visto anteriormente, o modelo ARIMA é recomendado para séries temporais não estacionárias, ou integradas. A Tabela [5] apresenta as melhores métricas atingidas para cada mercado e qual técnica de autoregressão foi utilizada. Também é possível perceber que o mercado da América do Norte foi o único que não obteve um r^2 positivo. Isso indica que apesar de todas as técnicas e variações nos modelos aplicadas, não se conseguiu encontrar uma configuração que se adequasse bem aos dados deste mercado.

Mercado	r^2	MAE	Técnica
AN	-0,1484	5075,3100	IMA
AS	0,0188	18327,6921	Autoreg
EU	0,2448	12070,6847	ARMA
AL	0,0381	987,7036	Autoreg
JP	0,2569	8693,4930	ARMA
BRL	0,2338	368,0827	XGBoostRegressor

Tabela [5] - Resultados das melhores técnicas de auto regressão, para cada mercado.

Os melhores resultados obtidos foram dos mercados EU e JP, e o mercado BRL, que apesar de não ser estacionário, apresenta uma variação pequena em seus valores ao longo da série, o que pode ter ajudado os modelos no treino e teste dos dados. Contudo, ao analisarmos o MAE, apenas o modelo aplicado sobre os mercados BRL e AL apresentaram erros considerados baixos, de cerca de 368 e 1059 unidades, respectivamente, para uma estimativa real satisfatória. Apesar destes resultados, um r^2 abaixo de 0,6 não pode ser considerado satisfatório para a previsão de valores.

Por fim, os gráficos 11 a 16 apresentam as estimativas para cada um dos seis mercados estados. O Gráfico [14] apresenta os valores estimados até 2017, ou seja, incluindo valores de 2016 para o mercado AL, e os Gráficos Gráfico [11], Gráfico [12], Gráfico [13], Gráfico [15] e Gráfico [16] apresentam os valores estimados até 2023, isto é, incluindo-se os valores de 2022, para os demais mercados.

O modelo do prophet previu os valores baseando-se nas tendências mais recentes dos volumes de vendas. Para o mercado da América do Norte e da Ásia, ele manteve uma previsão de crescimento, como visto nos Gráfico [11] e Gráfico [12]. Para o mercado da Europa, que apresentou uma aparente queda no último ano, o modelo estimou uma continuação deste declínio para 2022, como observado no Gráfico [13].

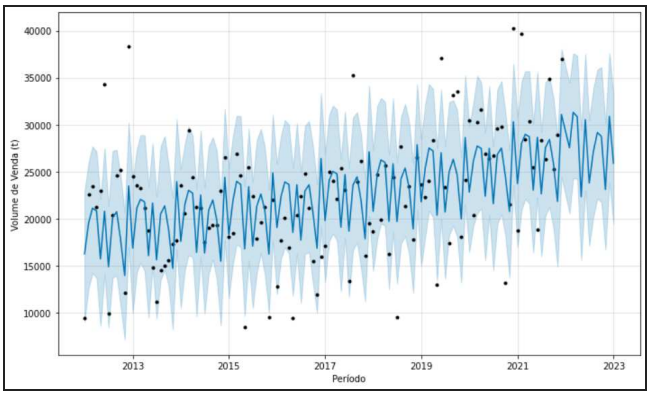

Gráfico [11]- Estimativa do modelo *Prophet* para o mercado AN, incluindo os valores para 2022.

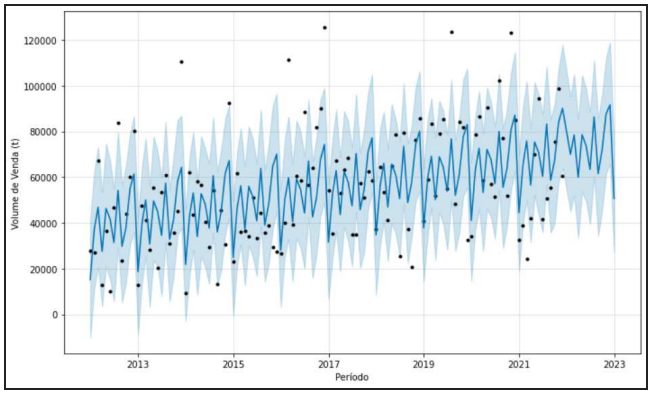

Gráfico [12] - Estimativa do modelo *Prophet* para o mercado AS, incluindo os valores para 2022.

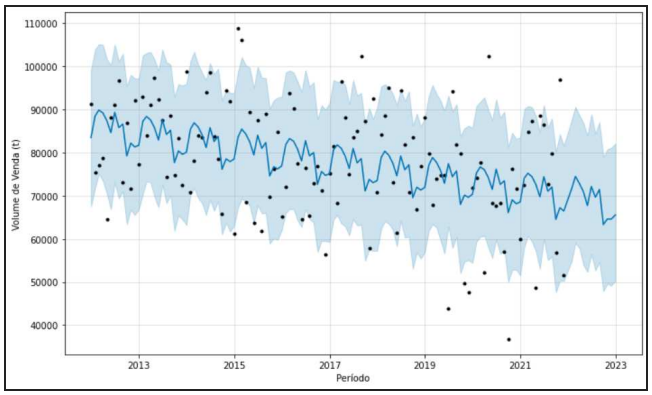

Gráfico [13] - Estimativa do modelo *Prophet* para o mercado EU, incluindo os valores para 2022.

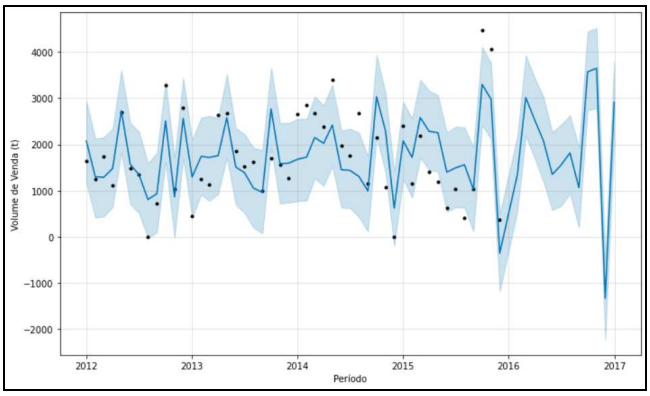

Gráfico [14] - Estimativa do modelo *Prophet* para o mercado AL, incluindo os valores para 2022.

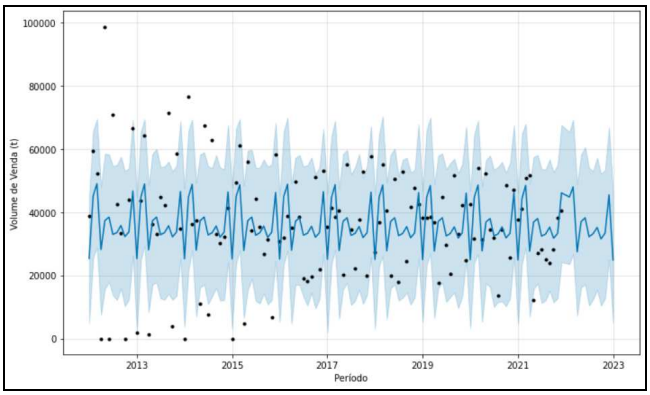

Gráfico [15] - Estimativa do modelo *Prophet* para o mercado JP, incluindo os valores para 2022.

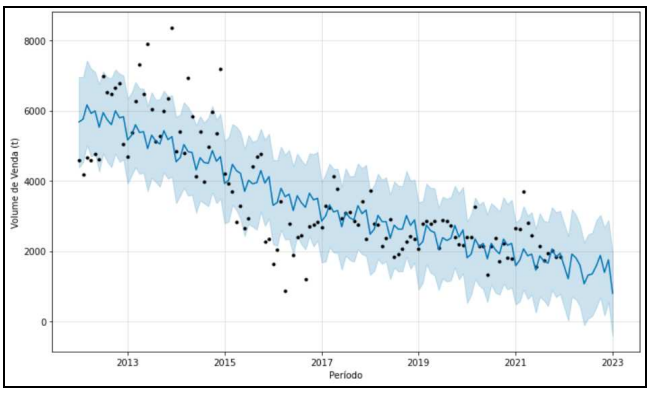

Gráfico [16] - Estimativa do modelo *Prophet* para o mercado BRL, incluindo os valores para 2022.

O mercado AL, por possuir poucos dados, mostrou-se difícil de prever, apresentando um comportamento anômalo, como visto no gráfico 18. Para o mercado JP, previu-se uma constância nos valores, segundo o gráfico 19. Para o mercado BRL, como ele apresentou uma leve queda nos últimos anos, o modelo estimou a manutenção desta queda, conforme esperado.

4.5 Ameaças para a validade

4.5.1 Ameaças à validade de conclusão

Apesar da grande quantidade de dados na base de dados originais, o total de *Volume de Vendas de* cada mercado estudado foi de apenas 120 ocorrências. É possível que esta baixa quantidade tenha afetado os testes de estacionariedade e as predições dos modelos. Uma maior base, com mais valores históricos, poderia melhorar o desempenho dos modelos.

4.5.2 Ameaças à validade interna

Devido à natureza dos dados estudados, que se caracterizam por serem séries temporais, a sequência e o período em que os dados foram coletados são muito importantes. Desta forma, dados coletados em diferentes períodos e com sequências diferentes, como por exemplo em vez de serem mensais, serem trimestrais, podem afetar a validade interna do estudo. Outra ameaça possível seria a utilização de diferentes bibliotecas dos modelos. Existem outras linguagens de programação, com outras bibliotecas, mas que aplicam os mesmos modelos estudados. Contudo, as particularidades de cada linguagem e a forma como estas outras bibliotecas foram construídas também podem afetar o resultado final.

4.5.3 Ameaças à validade de Constructo

As ameaças de constructo para este estudo são de natureza de design. É possível que nem todas as propriedades dos dados analisados tenham sido observadas. A aplicação de outras técnicas de análise de estacionariedade e de sazonalidade poderiam apresentar resultados diferentes, por exemplo.

4.5.4 Ameaças à validade externa

Devido à natureza do tipo de dado utilizado neste estudo, caso os mesmos tratamento seja utilizado em outra base, em uma possível extrapolação de cenário, poderia apresentar resultados diferentes.

5. TRABALHOS RELACIONADOS

Os conceitos de Econometria continuam sendo bastante aplicados atualmente. Existem diversos trabalhos disponíveis associando regressão, séries temporais e algum tipo de indicador. No estudo de Pemathilake *et al.* [22] foi realizada uma comparação entre os modelos ARIMA e uma Rede Neural Recorrente Híbrida (ARIMA com uma Rede Neural Recorrente - RNN) na previsão de vendas. Estes modelos foram avaliados usando-se a Média de Porcentagem de Erro Absoluta (do inglês, *Mean Absolute Percentage Error* - MAPE). De modo geral, os modelos ARIMA e RNN, quando aplicados sozinhos, tiveram um desempenho fraco se comparados ao modelo híbrido da Rede Neural recorrente em conjunto com o modelo ARIMA. O presente trabalho realizou um estudo do ARIMA e outros modelos, mas individualmente. Além disso, optou-se pela Média de Erro Absoluta (MAE), para se avaliar melhor a dimensão do valor do erro.

O trabalho realizado por Panay *et al.* [21] apresenta a previsão de indicadores de lojas de venda usando regressão interpretável. Neste estudo, foram analisados 3 indicadores: movimento de pessoas, taxa de conversão de compra e total de vendas durante um certo período de tempo. Foi aplicado o modelo de Regressão de Evidência Ponderada (do inglês, *Weighted Evidence Regression Model* - WEVREG) e as métricas para

avaliar sua acurácia foram RMSE (do inglês, *Rooted Mean Squared Error*) e MAPE. Ao fim, seus resultados mostraram um nível de precisão semelhante ao de outros métodos utilizados na literatura.

Um outro estudo, realizado por Silva [28] e aplicado em uma Universidade do Brasil, envolveu a aplicação do modelo ARIMA para previsão da demanda interna (consumo) no almoxarifado da Universidade Federal de Santa Maria (UFSM). Para a execução dos testes dos parâmetros (p,d,q) do modelo ARIMA foram seguidos os procedimentos do método Box-Jenkins. Para a avaliação do desempenho do modelo utilizou-se a métrica MAPE. Dos produtos presentes no almoxarifado, analisou-se apenas o consumo do papel higiênico, pois foi o único que apresentou dados mais estruturados. Foram realizadas análises e testes de estacionariedade, por meio dos correlogramas e do teste de raiz unitária de Kwiatkowski-Phillips-Schmidt-Shin (KPSS). No presente trabalho, um teste de estacionariedade também foi realizado, chamado de *adfuller*. A série temporal estudada mostrou-se estacionária. O estudo identificou cinco variações nos parâmetros (p,d,q) com resultados significativos.

Por fim, Novanda *et al*. [17] realizou uma comparação de diferentes técnicas de previsão, mas aplicadas aos preços do café em seu artigo. Os preços de café analisados foram adquiridos do Index Mundial de preços de café e do Centro de Estatística e Ministério da Agricultura da Indonésia. Os preços variaram mensalmente, de Janeiro de 2008 a Dezembro de 2016. Os modelos testados neste estudo foram Média Móvel (MA), ARIMA e de decomposição. O modelo de MA foi avaliado pela métrica MAPE; o modelo ARIMA pelo coeficiente de determinação (r^2) e o modelo de decomposição foi avaliado com as métricas MAE e MSD (do inglês, *Mean Squared Deviation*). Após as análises da métricas, concluiu-se que o melhor modelo para prever estes dados foi o ARIMA.

6. CONCLUSÃO

Os indicadores de vendas são um dos mais utilizados para avaliação dos negócios. O *Volume de Vendas* é um indicador fácil de ser coletado, mas sua análise e predição é algo bem mais complexo. A análise de regressão com apenas apenas o valores da série temporal do *Volume de Vendas* não se mostraram suficientes para se obter bons resultados. A natureza dos dados destas empresas se mostraram extremamente aleatórios, com grandes diferenças e particularidades para cada mercado. Percebe-se que a aplicação de uma técnica teve bons resultados para um mercado, mas para outros trouxe resultados muito ruins, não tornando possível uma generalização de um modelo único para se analisar estes dados de uma vez. Além disso, o fornecimento dos dados também se mostrou incompleto. Haviam muitos dados faltantes, a empresa que forneceu os dados não tinha muitas outras informações, como por exemplo, a composição dos mercados América do Norte e América Latina. Nenhum outro atributo também estava presente no conjunto inicial de dados, além do valor de ocorrência *(Volume de Vendas*). A existência de outros atributos, como preço médio do mercado ou tamanho da população de cada mercado, poderiam ajudar na utilização de outros modelos preditivos, que poderiam vir a ser mais assertivos do que a auto regressão.

7. AGRADECIMENTOS

Agradeço à Deus por tornar possível minha chegada a esta etapa final do curso, mesmo depois de dois anos muito difíceis para todo o mundo; à minha família e à minha esposa, Hingrid Medeiros, pelo todo apoio e motivação, mesmo distante muitos quilômetros nos últimos anos; ao meu irmão e companheiro de curso, Danilo Medeiros, que ao longo de quase cinco anos sempre me ajudou a superar as dificuldades do curso; aos amigos que fiz durante os últimos semestres nos projetos de Pesquisa e Desenvolvimento e à professora Melina Mongiovi, por sua paciência e por compartilhar seu conhecimento nesta jornada.

8. REFERÊNCIAS

[1] CHARNET, R. FREIRE, C. A. L. CHARNET, E. M. R. BONVINO, H. **Análise de Modelos de Regressão Linear**: com aplicações. 2ª Edição. Campinas, SP: Editora Unicamp, 2015.

[2] CHU, X. *et al*. **Data Cleaning**: Overview and Emerging Challenges. San Francisco, CA. 2016.

[3] DALAL, M. LI, A. C. TAORI, R. **Autoregressive Models**: What Are They Good For?. Berkeley, 2019.

[4] EHLERS, R.S. **Análise de Séries Temporais**. 4ª Edição. UFPR, Paraná: Laboratório de Estatística e Geoinformação. Paraná, 2007.

[5] FUNDAÇÃO NACIONAL DA QUALIDADE. **Sistema de Indicadores**.2014. Disponível em: https://fnq.org.br/comunidade/e-book-3-sistema-de-indicadores/. Acesso em: 11 ago. 2022.

[6] GÉRON, A. **Mãos à Obra:** Aprendizado de Máquina com Scikit-Learn & Tensorflow. Rio de Janeiro: Editora Alta Books, 2019.

[7] GU, R. S. **Data Cleaning Framework**: An Extensible Approach to Data Cleaning. Illinois, 2010.

[8] GUJARATI,D.N. PORTER, D.C. **Econometria Básica**. 5ª Edição.Nova York, NY: Editora AMGH, 2011.

[9] ICO. **International Coffee Organization**. Londres, 2011.

[10] JUPYTER. **Project Jupyter**. Disponível em: https://jupyter.org/. Acesso em: 11 ago. 2022.

[11] KAGANSKI, S *et al*. **Implementation of key performance indicators selection model as part of the Enterprise Analysis Model**. Estonia, 2017.

[12] MATÉ, A. TRUJILLO, J. MYLOPOULOS, J. **Conceptualizing and Specifying Key Performance Indicators in Business Strategy Models**. Espanha, 2012.

[13] MATPLOTLIB. **Matplotlib**: Visualization with Python. Disponível em:https://matplotlib.org/. Acesso em: 07 ago. 2022.

[14] MORETTIN, P.A. and TOLOI, C.M. **Previsão de Séries Temporais**. 2ª Edição. São Paulo: Editora Atual, 1987.

[15] MÜLLER, H. FREYTAG, J.C. **Problems, Methods, and Challenges in Comprehensive Data Cleansing**. Berlim, 2003.

[16] NICA, I. CHIRITA, N. IONESCU, S. **Using of KPIs and Dashboard in the analysis of Nike company's performance management**. Romania, 2021.

[17] NOVANDA, R. R. *et al*. **A Comparison of Various Forecasting Techniques for Coffee Prices**. Indonesia,2018.

[18] NUMPY. **Numpy**. Disponível em: https://numpy.org/. Acesso em: 06 ago. 2022..

[19] SILVA. L. **Análise da Aplicação do Modelo ARIMA**: estudo em uma instituição federal de ensino superior. Santa

Maria, RS, 2017.
[20] PANDAS. **Pandas documentation**. Disponível em: https://pandas.pydata.org/pandas-docs/stable/index.html#**.** Acesso em: 08 ago. 2022.
[21] PANAY, B *et al*. **Forecasting Key Retail Performance Indicators Using Interpretable Regression**. Chile, 2021.
[22] PEMATHILAKE, R. G. H. *et al*. **Sales Forecasting Based on AutoRegressive Integrated Moving Average and Recurrent Neural Network Hybrid Model**. Sri Lanka, 2018.
[23] PMDARIMA. **Pmdarima:** estimators for python.Disponível em: http://alkaline-ml.com/pmdarima/. Acesso em: 08 ago. 2022.
[24] PROPHET. **Prophet**: forecasting at scale. Disponível em: https://facebook.github.io/prophet/. Acesso em: 08 ago. 2022.
[25] RIDZUAN, F. ZAINON, W. M. N. W. **A Review on Data Cleansing Methods for Big Data**. Malásia, 2019.
[26] SCIKIT-LEARN. **Scikit-learn**: machine learning in python. Disponível em: https://scikit-learn.org/stable/index.html. Acesso em: 06 ago. 2022.
[27] SEABORN. **Seaborn**: Statistical data visualization. Disponível em: https://seaborn.pydata.org/. Acesso em: 06 ago. 2022.
[28] SILVA, L. Rabenschlag, D. R. **Análise da Aplicação do Modelo Arima:** estudo em uma instituição federal de ensino superior. Universidade Federal de Santa Maria, Santa Maria. 2017.
[29] STATSMODEL. **Statsmodel** .Disponível em: https://www.statsmodels.org/stable/index.html. Acesso em: 07 ago. 2022.
[30] XGBOOST. **XGBoost Documentation**. Disponível em: https://xgboost.readthedocs.io/en/stable/index.html. Acesso em: 21 ago. 2022.
[31] WOOLDRIDGE, J. M. **Introductory Econometrics**.5ª Edição. Mason, OH: Editora Cengage Learning,2013.