



UNIVERSIDADE FEDERAL FLUMINENSE
CURSO DE ENGENHARIA DE PRODUÇÃO

DAÍSE CABRAL DE SOUZA
JÉSSICA CRISTINA DA COSTA MARTINS
JOÃO VICTOR RESNIK ORLANDINI

PREVISÃO DE DEMANDA DE BLUSAS BÁSICAS MASCULINA PARA A
EMPRESA SEGROB NOTLAD

ANÁLISE PREDITIVA

PROFESSOR DALTON BORGES

RIO DAS OSTRAS

2025

RESUMO

A proposta do presente estudo se baseia em um desafio real de previsão de demanda promovido pela marca brasileira de fast fashion Segrob Notlad, cujo objetivo principal é estimar o volume de camisetas básicas vendidas diariamente durante o mês de dezembro de 2024. Com forte presença no mercado nacional e expansão internacional, a Segrob Notlad é reconhecida por sua abordagem ousada, inovação constante e forte conexão com o público jovem urbano. A empresa tem investido em inteligência de mercado e automação como pilares de sua estratégia de crescimento.

O tema central da iniciativa é a aplicação da ciência de dados para previsão de demanda no setor de moda. O objetivo do desafio é testar a capacidade de análise preditiva de participantes, utilizando um banco de dados histórico que abrange o período de janeiro de 2022 a novembro de 2024. A proposta insere-se no contexto da transformação digital vivida pela empresa, que visa tornar seu planejamento mais ágil e responsivo às oscilações do mercado.

A metodologia sugerida envolve o uso de técnicas de modelagem estatística e algoritmos de inteligência artificial, como séries temporais, regressão e modelos de machine learning, além da incorporação de variáveis externas que possam impactar a demanda (ex: sazonalidade, eventos, campanhas publicitárias). O desafio também simula o ambiente empresarial real ao permitir mudanças de escopo e introdução de novas variáveis durante o processo.

Embora os resultados ainda estejam em construção, espera-se que os modelos gerados possibilitem maior assertividade nas decisões de produção e abastecimento, reduzindo perdas e otimizando o atendimento à demanda. Os participantes vencedores serão premiados com livros sobre ciência de dados e inteligência de negócios, incentivando o desenvolvimento contínuo de competências técnicas e estratégicas.

Diante do exposto, o desafio proposto pela Segrob Notlad representa uma iniciativa inovadora que aproxima a prática acadêmica da realidade do mercado. Ao integrar dados históricos com técnicas avançadas de previsão, a marca não apenas fortalece seu posicionamento como empresa orientada por dados, mas também

contribui para a formação de profissionais mais preparados para os desafios do varejo moderno.

Palavras-chave: Previsão de demanda; Inteligência artificial; Fast fashion; Análise preditiva; Varejo.

SUMÁRIO

1. INTRODUÇÃO.....	6
2. METODOLOGIA - CRISP-DM.....	7
2.1 Entendimento do Negócio.....	7
2.2 Entendimento dos Dados.....	8
2.3 Preparação dos Dados.....	9
2.4 Modelagem.....	10
2.5 Avaliação.....	11
2.6 Implantação.....	13
3. DESENVOLVIMENTO.....	14
3.1 Entendimento do Negócio.....	14
3.2 Entendimento dos Dados.....	15
3.3 Preparação dos Dados.....	32
3.4 Modelagem.....	34
3.5 Modelos de IA.....	48
4. PREVISÃO DE DEMANDA.....	57
5. CONCLUSÃO.....	60
6. REFERÊNCIAS BIBLIOGRÁFICAS.....	61

1. INTRODUÇÃO

A Segrob Notlad é uma marca brasileira de *fast fashion* consolidada nos mercados nacional e internacional por unir design versátil, preços acessíveis e uma comunicação ousada voltada ao público jovem, urbano e conectado. Fundada no Rio de Janeiro, a empresa cresceu rapidamente e hoje conta com mais de 80 lojas no Brasil, unidades conceito na Europa e presença estratégica na América do Sul, consolidando-se como uma referência em moda acessível e inovadora.

A história da marca é marcada pela fusão entre raízes culturais e uma visão empreendedora arrojada. Seu fundador, Segrob Notlad, natural de Ainarres, na Croácia, trouxe ao Brasil um olhar que mescla a estética minimalista europeia com a energia vibrante do estilo carioca. Essa identidade se manifesta tanto nas coleções quanto na cultura organizacional da empresa, caracterizada por agilidade, ousadia e foco em estratégias orientadas por dados.

Atualmente, a Segrob Notlad adota uma postura data-driven, investindo fortemente em inteligência de mercado, automação logística e tecnologias preditivas para aumentar a eficiência da cadeia produtiva. Em um setor altamente dinâmico e sensível à sazonalidade — como o varejo de moda — a previsão de demanda se tornou um diferencial competitivo essencial. Antecipar oscilações de consumo em períodos críticos como Black Friday, Natal e ações promocionais permite reduzir perdas, otimizar estoques e alinhar produção e vendas de forma mais estratégica.

Neste cenário, o presente projeto de análise preditiva de demanda foi desenvolvido em parceria com a Universidade Federal Fluminense – ICT de Rio das Ostras. O objetivo central é prever, com base em dados históricos de vendas entre janeiro de 2022 e novembro de 2024, o volume diário esperado de vendas de camisetas básicas masculinas para o mês de dezembro de 2024 — um dos produtos mais estratégicos da marca.

Este relatório documenta todo o processo analítico, desde o entendimento do negócio até a modelagem e avaliação dos resultados. Mais do que resolver um desafio pontual, o trabalho visa contribuir para a transformação digital da empresa, integrando moda, dados e inteligência de mercado em um único projeto.

2. METODOLOGIA - CRISP-DM

O CRISP-DM (Cross Industry Standard Process for Data Mining) é uma metodologia de mineração de dados em formato cíclico que é amplamente utilizada para o desenvolvimento de projetos de mineração de dados. Ele fornece uma estrutura organizada que orienta o processo desde a identificação do problema de negócio até a aplicação dos resultados em um ambiente de produção.

O processo é dividido em seis fases, que podem ser revisitadas conforme necessário.

2.1 Entendimento do Negócio

A fase de Entendimento do Negócio, conforme previsto na metodologia CRISP-DM, tem como finalidade garantir que o projeto analítico esteja alinhado com os objetivos estratégicos da organização. Para isso, é necessário compreender o setor de atuação da empresa, seu posicionamento competitivo e os desafios operacionais que podem ser mitigados com o uso de ciência de dados.

Neste projeto, essa etapa foi conduzida a partir da análise de documentos fornecidos pelo professor, visando mapear as necessidades reais do negócio. A partir desse processo, foi identificado como problema central a dificuldade da empresa em antecipar, com precisão, a demanda por um de seus produtos de maior giro: camisetas básicas masculinas. Este desafio se torna ainda mais crítico no mês de dezembro, período de alta sazonalidade e sensibilidade a eventos promocionais como a Black Friday e o Natal.

Com base nesse diagnóstico, o problema de negócio foi definido como a necessidade de prever a demanda diária por camisetas masculinas durante dezembro de 2024, de forma a subsidiar as áreas de produção, logística e marketing com informações confiáveis para tomada de decisão. Essa definição orienta todas as demais fases do projeto, garantindo que a análise preditiva desenvolvida gere valor real ao negócio e esteja alinhada às metas operacionais e comerciais da organização.

2.2 Entendimento dos Dados

Para compreender a fundo o comportamento da série temporal de vendas analisada, serão aplicadas diversas abordagens exploratórias e descritivas, com o objetivo de identificar padrões, variações e possíveis anomalias nos dados ao longo do tempo.

Inicialmente, será apresentada uma visão geral da série, que abrange um período de 1.065 dias consecutivos. Nessa etapa, serão destacados indicadores estatísticos fundamentais, que fornecem uma noção preliminar da amplitude e variabilidade dos dados ao longo do tempo.

Em seguida, será realizada a decomposição da série temporal, por meio de gráficos que separam os componentes principais: tendência e sazonalidade.

Será feita ainda uma análise segmentada da série por diferentes recortes de tempo, como ano, mês, semana do ano e semana do mês. Essa etapa busca evidenciar padrões que possam não estar visíveis na visão geral, como picos sazonais específicos em determinados períodos do ano ou mudanças de comportamento ao longo dos anos.

Para aprofundar a análise da distribuição dos dados, serão utilizados boxplots mensais e semanais, que permitem observar a mediana, a dispersão e a presença de outliers ao longo dos diferentes períodos. Essa visualização facilitará a comparação entre meses e semanas, evidenciando períodos mais voláteis ou com comportamento atípico.

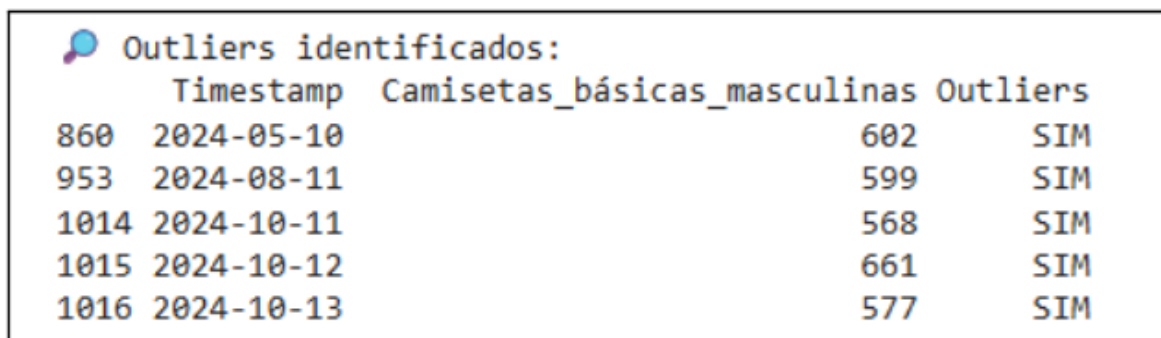
Por fim, será investigada a sazonalidade associada a feriados e campanhas promocionais, como a Black Friday. Esse estudo tem como objetivo identificar alterações no padrão de consumo em datas específicas, permitindo avaliar a influência desses eventos sobre o volume de vendas.

Essas abordagens combinadas fornecem uma base sólida para entender o comportamento da série temporal e para embasar futuras etapas do trabalho, como a modelagem preditiva e a tomada de decisão.

2.3 Preparação dos Dados

Antes de realizar as análises exploratórias e desenvolver modelos preditivos, é essencial aplicar uma etapa de preparação dos dados, com o objetivo de garantir a qualidade, a consistência e a relevância das informações utilizadas. Esta fase envolve três principais procedimentos metodológicos.

O primeiro passo consiste na identificação e tratamento de outliers. Valores atípicos podem distorcer as análises estatísticas e comprometer o desempenho dos modelos de previsão. Para isso, será utilizado o método do Intervalo Interquartil (IQR), que permite detectar pontos fora do padrão esperado com base na distribuição dos dados, nos dados obtivemos os seguintes outliers mostrados na Figura 1.



Outliers identificados:			
	Timestamp	Camisetas_básicas_masculinas	Outliers
860	2024-05-10	602	SIM
953	2024-08-11	599	SIM
1014	2024-10-11	568	SIM
1015	2024-10-12	661	SIM
1016	2024-10-13	577	SIM

Figura 1 - Outliers identificados.

Em seguida, será realizada a extração de variáveis temporais a partir da coluna de datas presente na série. Essa etapa envolve a criação de variáveis como dia da semana, número da semana, mês, ano, feriado e proximidade de datas comemorativas ou eventos promocionais. Essas variáveis adicionais permitem incorporar características temporais importantes na modelagem, enriquecendo a análise e contribuindo para a explicação de padrões sazonais e comportamentos específicos.

Por fim, serão criadas variáveis de defasagens (lags), que correspondem aos valores de vendas em dias anteriores. Neste estudo, serão considerados lags de 1 a

7 dias, permitindo capturar efeitos de dependência temporal de curto prazo. Essas variáveis são fundamentais para modelos que levam em consideração a autocorrelação presente em séries temporais, como modelos autoregressivos e híbridos.

Essa preparação cuidadosa dos dados é essencial para garantir resultados mais robustos e interpretações mais confiáveis nas etapas seguintes do trabalho.

2.4 Modelagem

Com os dados devidamente tratados e estruturados, serão aplicados diferentes modelos de previsão de séries temporais, com o objetivo de comparar abordagens e avaliar o desempenho preditivo de cada uma delas. Os modelos selecionados abrangem desde métodos simples e intuitivos até técnicas estatísticas mais sofisticadas, permitindo uma análise ampla e fundamentada. Serão utilizados os seguintes modelos:

- Naive (ingênuo): utiliza o valor mais recente como previsão para os períodos seguintes. É uma abordagem de referência que serve como base comparativa mínima.
- Seasonal Naive: extensão do modelo ingênuo que considera a repetição do valor de vendas no mesmo dia da semana anterior, capturando a sazonalidade semanal.
- Cumulativo (média histórica): utiliza a média acumulada de todas as observações anteriores como valor preditivo, representando uma tendência global constante.
- Média Móvel (7, 14 e 30 dias): calcula a média dos valores observados nos últimos dias, suavizando variações de curto prazo e permitindo captar tendências locais.
- Suavização Exponencial: aplica pesos decrescentes aos valores passados, dando mais importância às observações recentes e proporcionando uma resposta adaptativa às mudanças.
- Regressão Linear: modela a relação entre as vendas e variáveis temporais, permitindo estimar a tendência geral da série com base em uma função linear dos dados.

- Regressão Dinâmica com variáveis exógenas: modelo mais complexo que incorpora variáveis externas, como defasagens de vendas (lags), feriados e promoções, permitindo uma previsão mais sensível ao contexto e aos eventos externos.

Para garantir a robustez da análise e testar a capacidade preditiva dos modelos em dados não vistos, será adotado um processo de validação temporal. As últimas quatro semanas de 2024 serão reservadas como um conjunto de validação final, sobre o qual os modelos serão avaliados. Essa estratégia permite simular uma situação real de previsão futura e comparar o desempenho dos métodos por meio de métricas de erro, como MAD, MAPE e RMSE.

Essa abordagem sistemática visa não apenas identificar o modelo mais eficaz para o conjunto de dados analisado, mas também compreender as limitações e vantagens de cada técnica em contextos distintos.

2.5 Avaliação

A etapa de avaliação é fundamental para mensurar a qualidade das previsões geradas pelos modelos aplicados, permitindo identificar quais abordagens oferecem maior precisão e confiabilidade. Para isso, serão utilizadas métricas quantitativas de erro, análises visuais comparativas e técnicas de validação temporal, com o objetivo de garantir a robustez e a generalização dos resultados.

Cinco métricas principais serão utilizadas para quantificar o desempenho preditivo dos modelos:

- MAD (Mean Absolute Deviation): mede o erro médio absoluto entre os valores previstos e os valores reais, fornecendo uma noção direta da magnitude média dos desvios, sem penalizar fortemente os erros maiores.
- MAPE (Mean Absolute Percentage Error): expressa o erro médio absoluto em termos percentuais, o que facilita a interpretação em

relação ao volume total de vendas. É especialmente útil quando se deseja avaliar o impacto proporcional dos erros de previsão.

- RMSE (Root Mean Squared Error): calcula a raiz quadrada do erro quadrático médio. Por elevar os erros ao quadrado antes de somá-los, essa métrica penaliza mais fortemente os desvios extremos, sendo mais sensível a picos de erro.
- MAE (Mean Absolute Error): representa a média dos valores absolutos das diferenças entre as previsões e os valores observados. Assim como o MAD, fornece uma visão direta da precisão média do modelo, mas é mais comumente usada em contextos estatísticos e de machine learning. no contexto de previsão de dados, MAD e MAE são equivalentes
- MSE (Mean Squared Error): obtido ao elevar ao quadrado a diferença entre os valores previstos e reais e, em seguida, calcular a média. Diferente do MAE, o MSE aumenta o peso de grandes erros, tornando-o eficaz na identificação de modelos que falham em capturar picos ou quedas acentuadas.

Os resultados dessas métricas serão organizados em uma tabela comparativa, permitindo visualizar o desempenho relativo de cada modelo de forma objetiva. Além disso, serão apresentados gráficos que comparam as previsões com os valores reais, facilitando a identificação de padrões, desvios sistemáticos e possíveis falhas dos modelos em determinados períodos.

Com base nessas análises, será feita a escolha do modelo final, considerando tanto os resultados quantitativos quanto os aspectos interpretativos.

Para complementar a avaliação pontual no conjunto de validação final, será aplicada uma validação cruzada temporal (time series cross-validation). Essa técnica consiste em dividir a série em múltiplos subconjuntos sequenciais, simulando várias janelas de treinamento e teste ao longo do tempo.

Esse procedimento permite avaliar a robustez e estabilidade dos modelos em diferentes períodos históricos, assegurando que o bom desempenho não esteja restrito a um intervalo específico. Dessa forma, é possível verificar se o modelo mantém sua acurácia de forma consistente, mesmo diante de variações no comportamento da série.

A combinação de métricas de erro, comparações visuais e validação cruzada oferece uma base sólida para a seleção do modelo mais adequado e para a confiabilidade das previsões futuras.

2.6 Implantação

A etapa de implantação visa demonstrar a aplicação prática dos modelos preditivos desenvolvidos, com foco na utilização de algoritmos que possam ser incorporados a sistemas de apoio à decisão, automatizando a previsão de vendas e permitindo ações mais ágeis e assertivas por parte da gestão.

Nesta fase, será implementado o modelo discriminativo de Árvore de Decisão. Este tipo de modelo é capaz de aprender padrões de decisão a partir dos dados históricos, identificando combinações de variáveis que levam a determinados níveis de venda. Por sua natureza interpretável e eficiente, especialmente em conjuntos de dados tabulares, esse modelo é adequado para uma primeira aplicação prática, podendo ser facilmente integrado a sistemas operacionais e dashboards.

Além disso, serão discutidas as possibilidades de evolução futura do projeto, especialmente com a adoção de modelos generativos baseados em redes neurais recorrentes (RNN), como as Long Short-Term Memory (LSTM). Esses modelos são capazes de capturar dependências temporais de longo prazo e padrões complexos que modelos estatísticos tradicionais ou modelos de árvore de decisão podem não identificar plenamente. Embora requeiram maior capacidade computacional e volume de dados para treinamento, os modelos LSTM oferecem grande potencial para aplicações mais avançadas de previsão, especialmente em ambientes com alta variabilidade e múltiplos fatores influenciando as vendas.

A utilização atual de um modelo discriminativo com variáveis de lag representa uma solução eficaz e acessível, enquanto a possibilidade de adoção futura de modelos generativos abre caminho para análises mais profundas e previsões ainda mais precisas, conforme a maturidade dos dados e da infraestrutura analítica da organização evolua.

3. DESENVOLVIMENTO

3.1 Entendimento do Negócio

A Segrob Notlad é uma empresa brasileira do setor de *fast fashion*, caracterizada por sua atuação em um mercado altamente dinâmico e competitivo, com presença consolidada no território nacional e crescente inserção internacional. Seu modelo de negócios combina agilidade operacional, preços acessíveis e forte apelo estético voltado ao público jovem e urbano. Com mais de 80 lojas físicas e forte presença no comércio eletrônico, a empresa adota estratégias omnichannel para maximizar sua capilaridade digital e geográfica.

O mercado de moda varejista, especialmente no segmento *fast fashion*, demanda respostas rápidas às tendências, grande flexibilidade de produção e excelência logística. As principais características desse setor incluem:

- Alta rotatividade de produtos,
- Sazonalidade intensa,
- Mudança frequente de preferências do consumidor,
- Elevada sensibilidade a datas promocionais e feriados.

Nesse cenário, a previsão de demanda se torna um pilar estratégico. A incapacidade de prever adequadamente a demanda pode gerar estoques em excesso (com custos de armazenagem e liquidação) ou rupturas (perda de vendas e insatisfação do cliente). O desafio é ainda mais crítico em períodos como Black Friday, Natal e campanhas institucionais, quando a incerteza e o volume de vendas aumentam substancialmente.

Visando superar esses desafios, a Segrob Notlad tem investido em análise preditiva, automação da cadeia de suprimentos e tecnologias baseadas em IA. Seu objetivo é tornar o processo de tomada de decisão mais baseado em dados e menos sujeito a suposições, reduzindo perdas e otimizando os recursos ao longo da cadeia produtiva.

3.1.1 Problema de negócio

A empresa precisa antecipar com precisão a demanda por camisetas básicas masculinas, um de seus produtos mais estratégicos, no mês de dezembro de 2024, para garantir alinhamento entre produção, estoque e ações promocionais.

3.1.2 Problema analítico

Desenvolver um modelo preditivo com base em dados históricos (jan/2022 a nov/2024) que permita estimar as vendas diárias para dezembro de 2024, considerando variáveis sazonais, eventos externos e padrões de comportamento.

3.1.3 Objetivo do projeto

Apoiar o planejamento de produção, compras e distribuição da Segrob Notlad com uma previsão confiável e adaptada à natureza cíclica e promocional do mercado de moda. O projeto é desenvolvido em parceria com a Universidade Federal Fluminense – ICT Rio das Ostras, promovendo a aplicação acadêmica da ciência de dados em um caso real de mercado.

3.2 Entendimento dos Dados

Com base na base histórica de vendas diárias da Segrob Notlad, referente ao período de janeiro de 2022 a novembro de 2024, foi realizada uma análise exploratória aprofundada da série temporal, com o objetivo de identificar padrões sazonais, tendências estruturais, ciclos de consumo e comportamentos atípicos que possam subsidiar a modelagem preditiva da demanda para dezembro de 2024 mês estratégico para o setor varejista de moda.

3.2.1 Visão Geral da Série Temporal

A análise de séries temporais é uma etapa crítica dentro da metodologia CRISP-D. O entendimento da estrutura temporal das vendas permite identificar padrões repetitivos (sazonalidade), comportamentos sistemáticos de crescimento ou queda (tendência) e componentes aleatórios (resíduos) que devem ser considerados na modelagem preditiva.

Os gráficos a seguir representam a média de vendas por dia da semana nos anos de 2022, 2023 e 2024, e foram utilizados para avaliar o comportamento da série em frequência semanal, permitindo identificar componentes fundamentais para futuros modelos de regressão ou previsão de demanda.

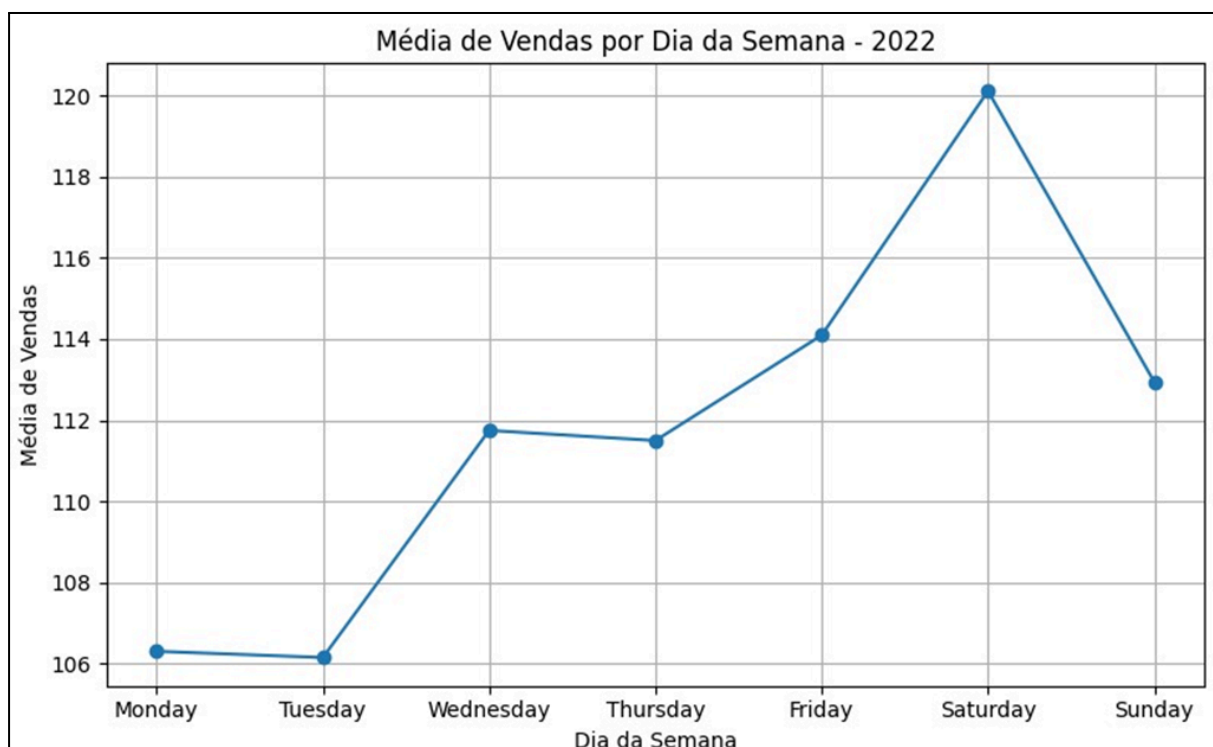


Figura 2 - Gráfico de Média de Vendas por Dia da Semana - 2022.

A série temporal semanal de 2022, ilustrada na Figura 2, revela um padrão sazonal semanal estável, com picos recorrentes aos sábados (acima de 120 unidades) e os menores valores concentrados entre segunda e terça-feira (~106 unidades). A série não demonstra uma tendência anual significativa nesse recorte, o que é coerente com uma fase inicial de estabilização de vendas.

O padrão é típico de ciclos de consumo orientados ao calendário, com sazonalidade de sete dias bem definida. A presença de resíduos (diferença entre o

valor observado e o valor ajustado pelo modelo) tende a ser baixa, mas deve ser observada nos finais de semana, devido à disparidade entre picos e vales.

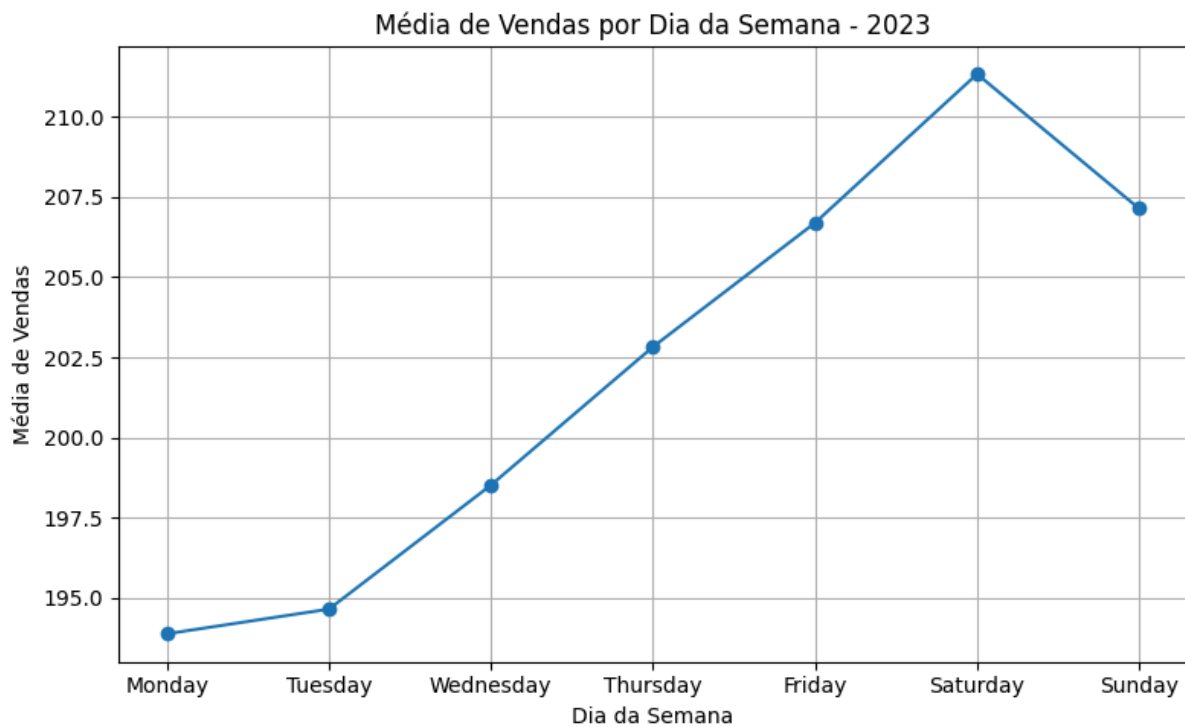


Figura 3 - Gráfico de Média de Vendas por Dia da Semana - 2023.

Em 2023, a série temporal semanal mantém a mesma estrutura sazonal, mas com valores médios significativamente mais altos, variando entre 194 (segunda-feira) e 211 (sábado). Esse aumento, que está sendo demonstrada na Figura 3, configura uma tendência positiva ao longo dos anos, indicando que o nível médio da série está se deslocando para cima.

A diferença entre os dias de menor e maior venda também aumentou, ampliando a amplitude do componente sazonal. A série, portanto, apresenta sinais de comportamento aditivo, onde tendência e sazonalidade se somam ao longo do tempo. Os resíduos devem ser monitorados com atenção crescente, pois a elevação na variabilidade semanal implica maior risco de erro para modelos que desconsiderem esse padrão.

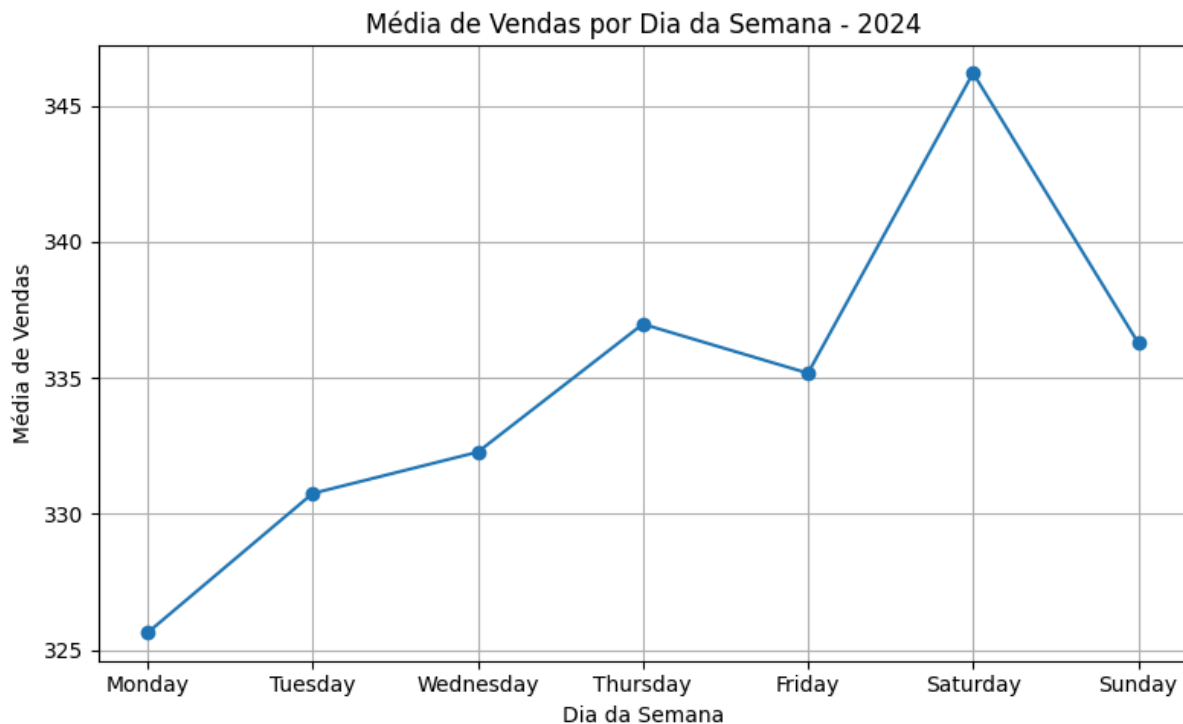


Figura 4 - Gráfico de Média de Vendas por Dia da Semana - 2024.

Em 2024, a série atingiu seu nível mais alto, com valores médios entre 325 (segunda-feira) e 346 (sábado). O padrão sazonal continua estável, reforçando a forte periodicidade semanal, mas agora com uma tendência de longo prazo claramente estabelecida. Esse cenário, apresentado na Figura 4, demonstra que a série possui os três componentes principais:

- Tendência crescente (elevação contínua nas vendas médias),
- Sazonalidade semanal estável (padrões cíclicos regulares a cada 7 dias),
- Resíduos moderados, provavelmente estruturados em torno de extremos (ex. picos de sábado).

Modelos preditivos que ignorarem esses componentes correm o risco de subajuste (underfitting) e perda de precisão.

A análise das séries temporais semanais referentes ao período de 2022 a 2024 evidencia um cenário propício para a aplicação de modelos preditivos que

levem em conta múltiplos componentes estruturais, como tendência, sazonalidade e resíduos. A presença de uma tendência de crescimento clara e contínua, associada a uma sazonalidade semanal bem definida e recorrente, sugere que modelos mais sofisticados são necessários para capturar adequadamente a dinâmica dos dados.

Nesse contexto, destacam-se como abordagens recomendadas: os modelos sazonais aditivos ou multiplicativos, como o SARIMA, capazes de incorporar diretamente os padrões de repetição temporal; os modelos de regressão linear multivariada com variáveis dummy para os dias da semana e outras variáveis temporais, que possibilitam uma explicação mais detalhada das variações periódicas; e os modelos baseados em árvores de decisão, como Random Forest e XGBoost, especialmente eficazes quando utilizados com codificação de datas, permitindo capturar interações não lineares e padrões complexos nos dados.

É fundamental observar que, enquanto a componente sazonal demonstra-se estacionária — ou seja, sua estrutura se mantém constante ao longo do tempo —, a componente de tendência é não estacionária, exigindo tratamento específico, como diferenciação ou modelagem explícita, para garantir a eficácia dos modelos de previsão. A correta identificação e tratamento desses componentes são determinantes para melhorar a acurácia e a confiabilidade das estimativas em contextos de demanda.

3.2.2 Padrões Anuais e Mensais

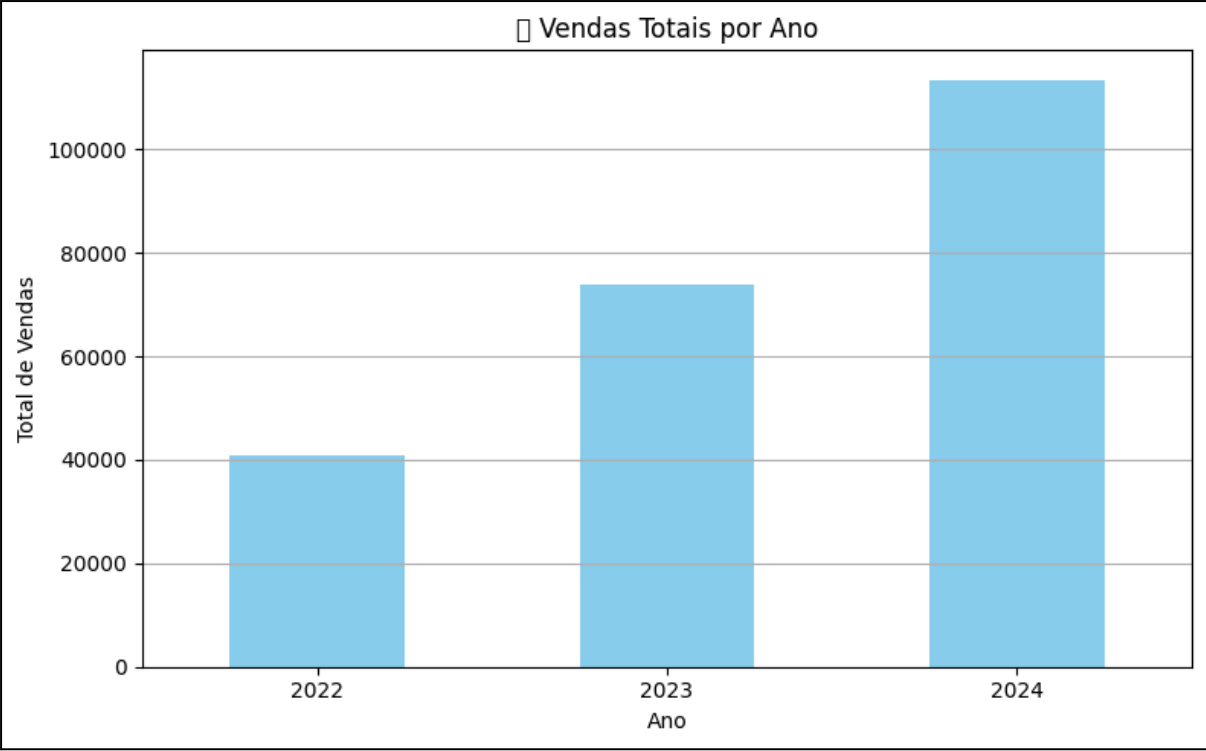


Figura 5 - Comparação de Vendas Totais por Ano.

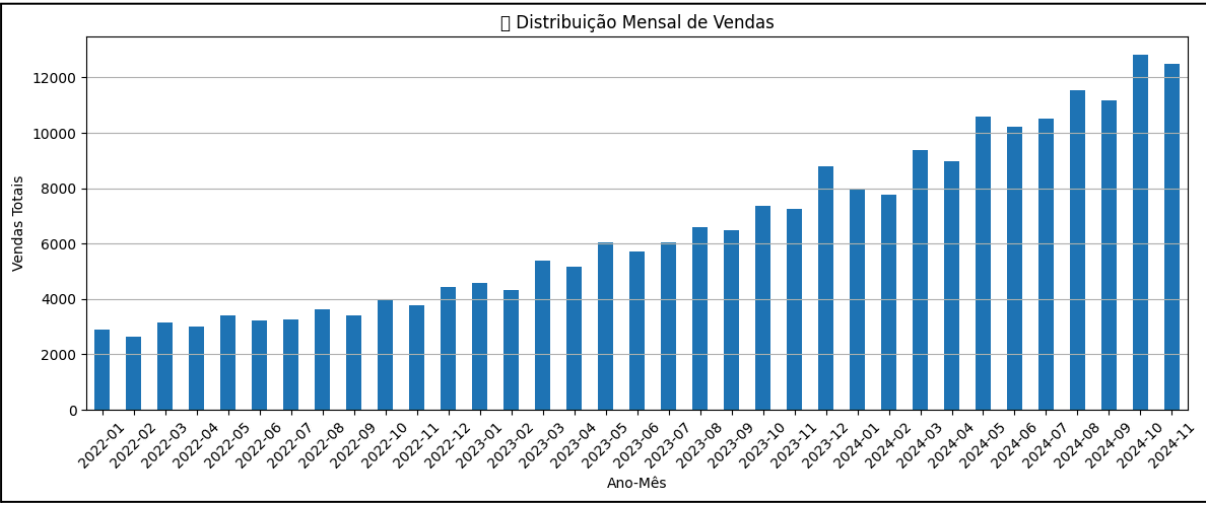


Figura 6 - Comparação de Vendas Totais por Mês.

A análise detalhada dos gráficos de vendas nas Figuras 5 e 6 revela uma tendência de crescimento robusta e consistente no desempenho da marca ao longo do período observado. Especificamente, o ano de 2024 se destaca com um volume de vendas visivelmente superior em comparação a 2022 e 2023, conforme ilustrado pela progressão ascendente nas vendas anuais. Mais notavelmente, o crescimento das vendas se acentua significativamente a partir do segundo semestre de 2024,

apresentando um impulso que é comum em períodos estratégicos do varejo, frequentemente impulsionado por eventos promocionais chave como Black Friday e festas de fim de ano.

Paralelamente, a média mensal de vendas demonstra uma progressão constante, um indicativo claro do amadurecimento comercial da marca, que não apenas consolida sua posição no mercado, mas também contribui para uma maior previsibilidade da demanda. Essa expansão ao longo de 2024 sugere que, além do crescimento orgânico, há uma forte probabilidade de que ações estratégicas como a intensificação de campanhas promocionais, a ampliação ou otimização de canais digitais e o lançamento de novas iniciativas institucionais tenham desempenhado um papel crucial.

A combinação desses fatores externos é fundamental para explicar o aumento na previsibilidade da demanda e, portanto, reforça a necessidade crítica de incorporar essas variáveis (por exemplo, como *dummy variables* para eventos ou métricas de marketing) na modelagem preditiva futura, a fim de construir previsões mais precisas e explicativas.

3.2.3 Identificação de Picos e Sazonalidade Estratégica

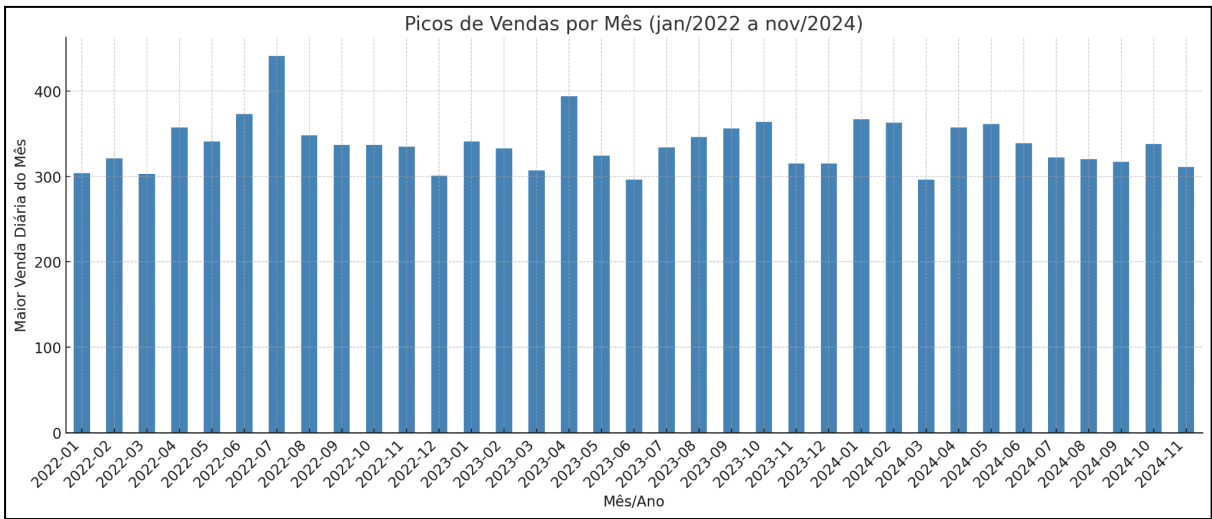


Figura 7 - Picos de vendas por Mês.

Conforme ilustrado na Figura 7, foram identificados picos recorrentes nos meses de agosto, setembro, outubro, novembro e dezembro, que se alinham a eventos do calendário promocional da empresa e do setor:

- Outubro: início da coleção primavera/verão;
- Novembro: Black Friday – pico absoluto de vendas;
- Dezembro: Natal e festas de fim de ano, com alta intenção de compra de vestuário;
- Agosto e Setembro: possível relação com campanhas institucionais ou aquecimento de vendas de meia-estação.

A análise indica que a distribuição estratégica de campanhas promocionais ao longo do segundo semestre tem contribuído para o aumento da intensidade da demanda, gerando padrões de consumo que não se repetem naturalmente, mas sim como resposta a estímulos de marketing. Esses ciclos artificiais criados pela empresa precisam ser devidamente reconhecidos e incorporados aos modelos preditivos, a fim de garantir maior aderência às flutuações reais do mercado.

3.2.4 Análise Semanal e Intra-Mensal

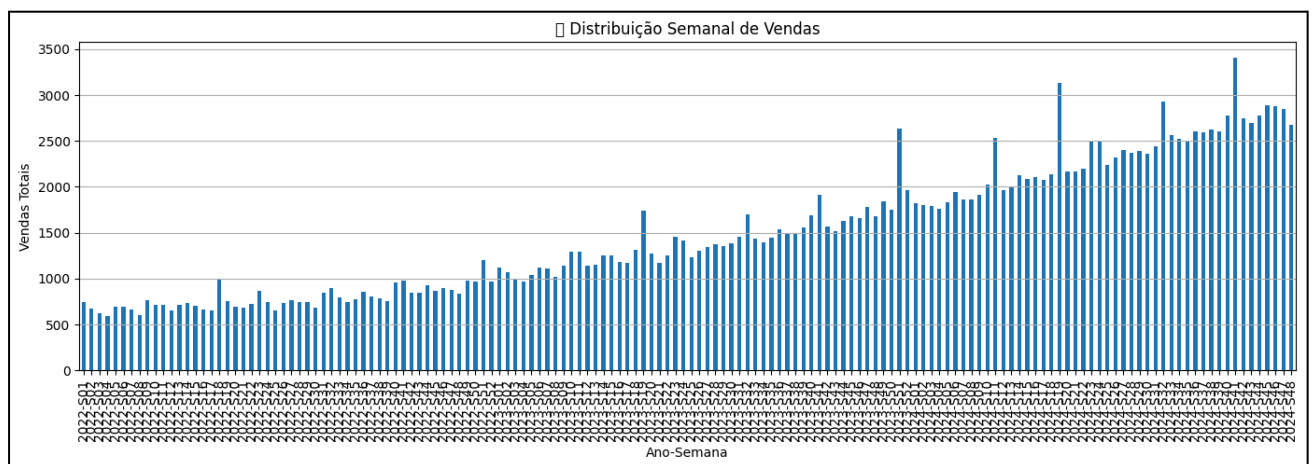


Figura 8 - Distribuição Semanal de Vendas.

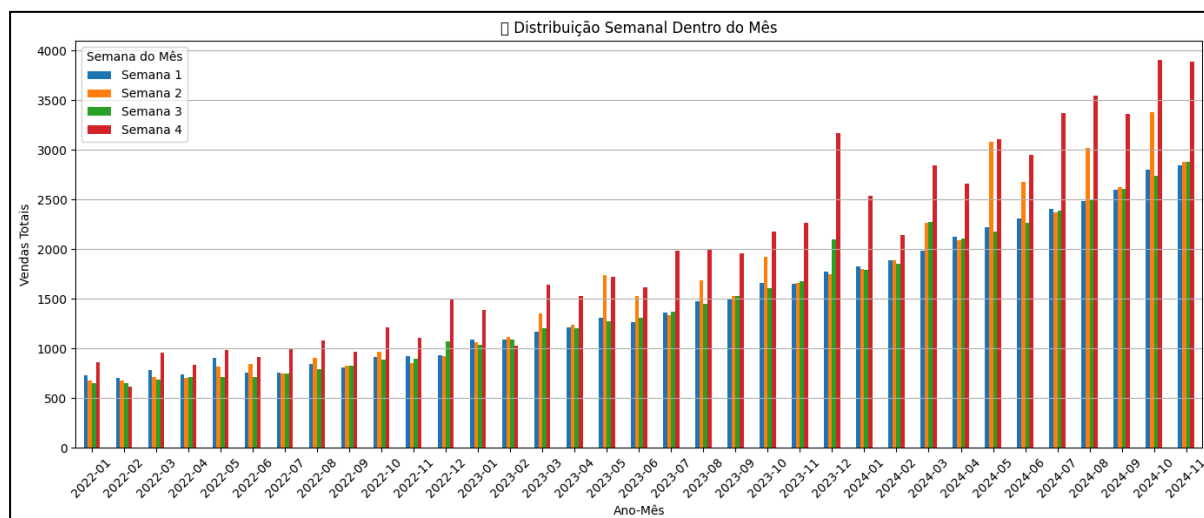


Figura 9 - Distribuição Semanal de Vendas Agrupada por Mês.

A análise aprofundada dos dados de vendas, suportada por uma decomposição de série temporal e visualizações específicas, revela padrões cruciais para a compreensão e previsão da demanda em diferentes granularidades, podendo ser observadas nas Figuras 8 e 9.

3.2.4.1 Tendência e Desempenho Anual

A série histórica de vendas demonstra uma clara e robusta tendência de crescimento contínuo ao longo dos três anos observados, refletindo um amadurecimento comercial da marca e uma possível ampliação de sua base de clientes. A análise anual confirma essa evolução positiva, com o ano de 2024 apresentando um desempenho visivelmente superior em volume de vendas em comparação a 2022 e 2023. Notavelmente, esse crescimento se intensifica de forma acentuada a partir do segundo semestre de 2024, um comportamento que frequentemente coincide com períodos promocionais estratégicos e de alta demanda no varejo.

A progressão constante da média mensal de vendas ao longo do período corrobora a estabilização e a maior previsibilidade da demanda. Essa expansão ao longo de 2024 sugere que, além do crescimento orgânico, há uma forte probabilidade de que a intensificação de ações promocionais, a ampliação dos

canais digitais e o lançamento de novas campanhas institucionais tenham sido fatores cruciais. A combinação desses elementos é fundamental para explicar o aumento na previsibilidade da demanda e, portanto, reforça a necessidade crítica de incorporar essas variáveis externas (como *dummy variables* para eventos ou métricas de marketing) na modelagem preditiva futura, a fim de construir previsões mais precisas e explicativas.

3.2.4.2 Sazonalidade Recorrente

Sazonalidade Mensal/Trimestral (Picos Anuais/Semestrais): A análise dos dados em granularidade mensal e semanal ao longo do ano revela picos de vendas recorrentes e altamente significativos nos meses de agosto, setembro, outubro, novembro e dezembro. Esses picos alinham-se consistentemente com eventos estratégicos do calendário promocional da empresa e do setor, incluindo:

- Outubro: Início da coleção primavera/verão.
- Novembro: Black Friday, que se estabelece como o pico absoluto de vendas do ano.
- Dezembro: Natal e festas de fim de ano, impulsionando a alta intenção de compra de vestuário.
- Agosto e Setembro: Possível relação com campanhas institucionais ou o aquecimento das vendas de meia-estação. A distribuição estratégica de campanhas promocionais ao longo do segundo semestre tem contribuído decisivamente para o aumento da intensidade da demanda, gerando padrões de consumo que são uma resposta direta a estímulos de marketing, e não apenas ciclos naturais de consumo. Esses "ciclos artificiais" criados pela empresa precisam ser devidamente reconhecidos e incorporados aos modelos preditivos para garantir maior aderência às flutuações reais do mercado e otimizar o planejamento de recursos, complementando a sazonalidade intra-mensal.

Sazonalidade Semanal/Quinzenal (Padrões Intra-mês): Em uma granularidade ainda mais fina, a análise semanal dentro de cada mês (intra-mensal) revela padrões cíclicos bem definidos, com destaque notório para as semanas 2 e 4,

que frequentemente concentram os maiores volumes de venda. Esse padrão quinzenal é altamente provável de estar relacionado a datas de pagamento (geralmente no início e meio do mês) e a ações promocionais periódicas que a empresa estrategicamente alinha a esses momentos de maior poder de compra do consumidor.

Essa concentração cíclica da demanda ao longo do mês aponta para oportunidades significativas de otimização logística (gestão de estoque, dimensionamento de equipe, etc.) e de um planejamento mais eficaz das campanhas de marketing, que podem ser ajustadas para coincidir precisamente com os momentos de maior propensão à compra, maximizando o retorno sobre o investimento.

3.2.4.3 Resíduos e Fatores Anômalos

A análise dos resíduos, que representa a parcela das vendas não explicada pela tendência e sazonalidade modeladas, revela uma distribuição predominantemente aleatória e centrada em zero. Embora isso seja um indicativo positivo de que as principais componentes da série temporal foram capturadas, a identificação de picos e vales pontuais (outliers) nos resíduos é crucial. Esses desvios sugerem a ocorrência de eventos não modelados, como promoções de curto prazo não rotineiras, falhas operacionais, ou outras influências externas e anômalas que impactaram as vendas em momentos específicos. A investigação dessas datas e a correlação com dados externos são essenciais para compreender e, se relevantes, incorporar essas anomalias em futuras análises e modelos preditivos, transformando ruído em informação útil.

3.2.5 Análise de Distribuição Estatística

A análise da distribuição estatística das vendas por diferentes recortes (mensal e semanal), utilizando visualizações como box plots e histogramas, ilustrada a seguir nas figuras 10, 11, 12 e 13, demonstra uma estrutura de dados geralmente estável, com características favoráveis à aplicação de modelos preditivos avançados

A baixa presença de outliers Extremos: As distribuições revelam uma baixa incidência de outliers extremos, ou seja, poucos pontos de dados que se desviam significativamente do padrão geral. Essa característica é crucial, pois favorece o uso de modelos estatísticos clássicos que são menos robustos a valores atípicos. A estabilidade resultante garante maior confiabilidade e estabilidade nas estimativas dos modelos preditivos.

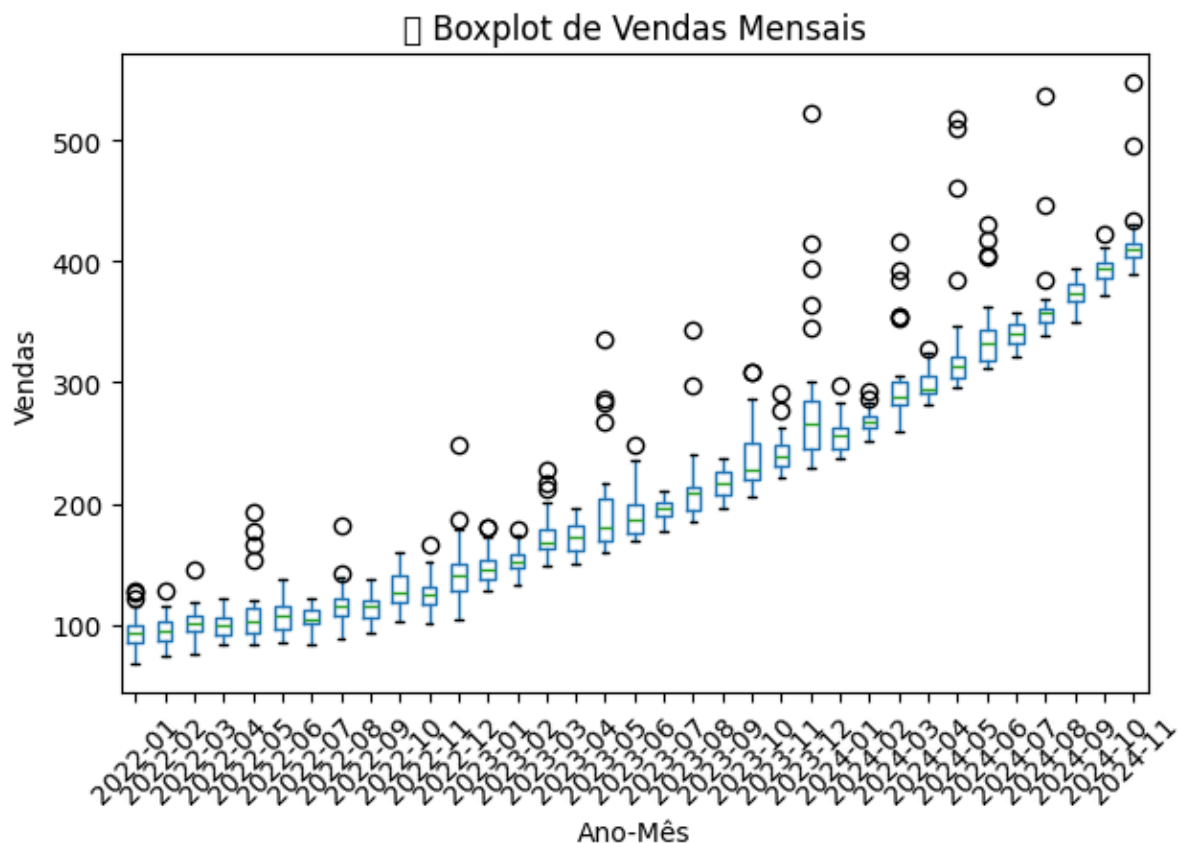


Figura 10 - Boxplot de Vendas Mensais.

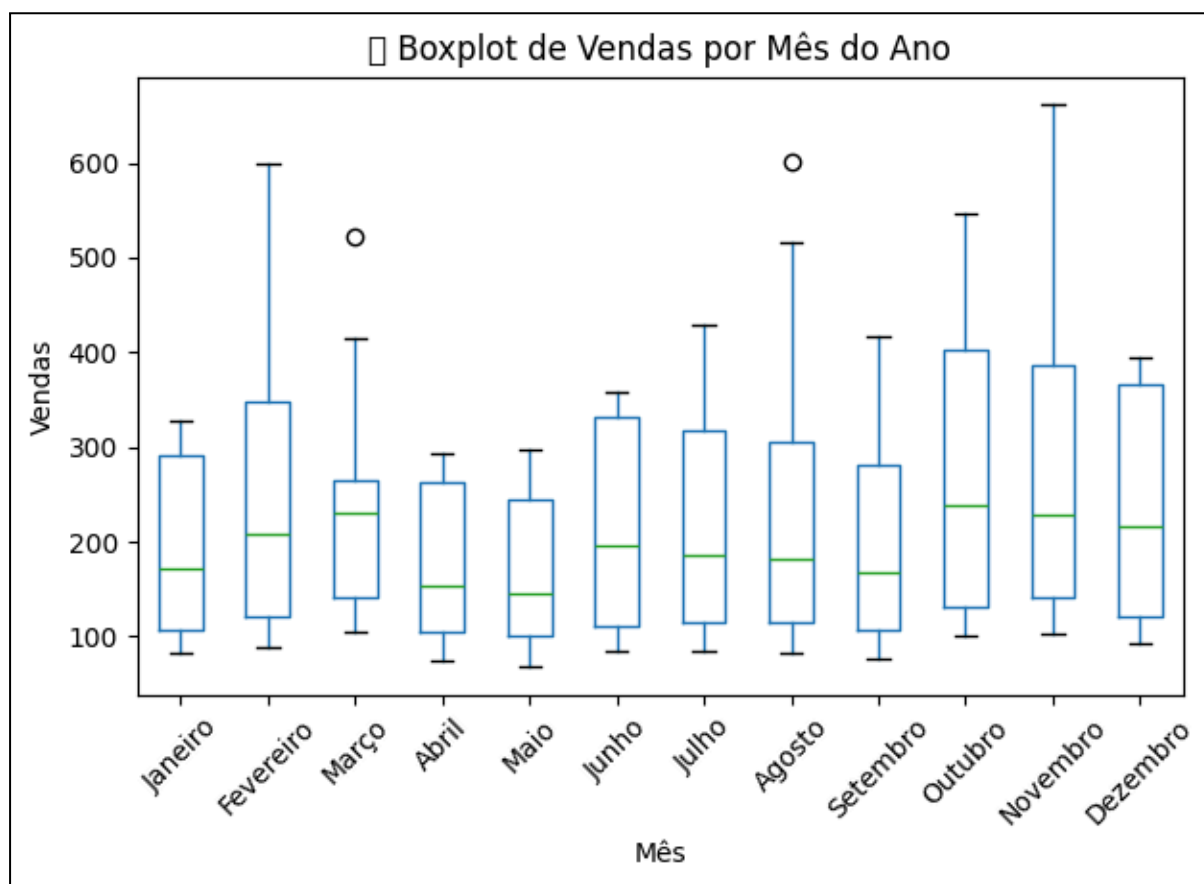


Figura 11 - Boxplot de Vendas por Mês do Ano.

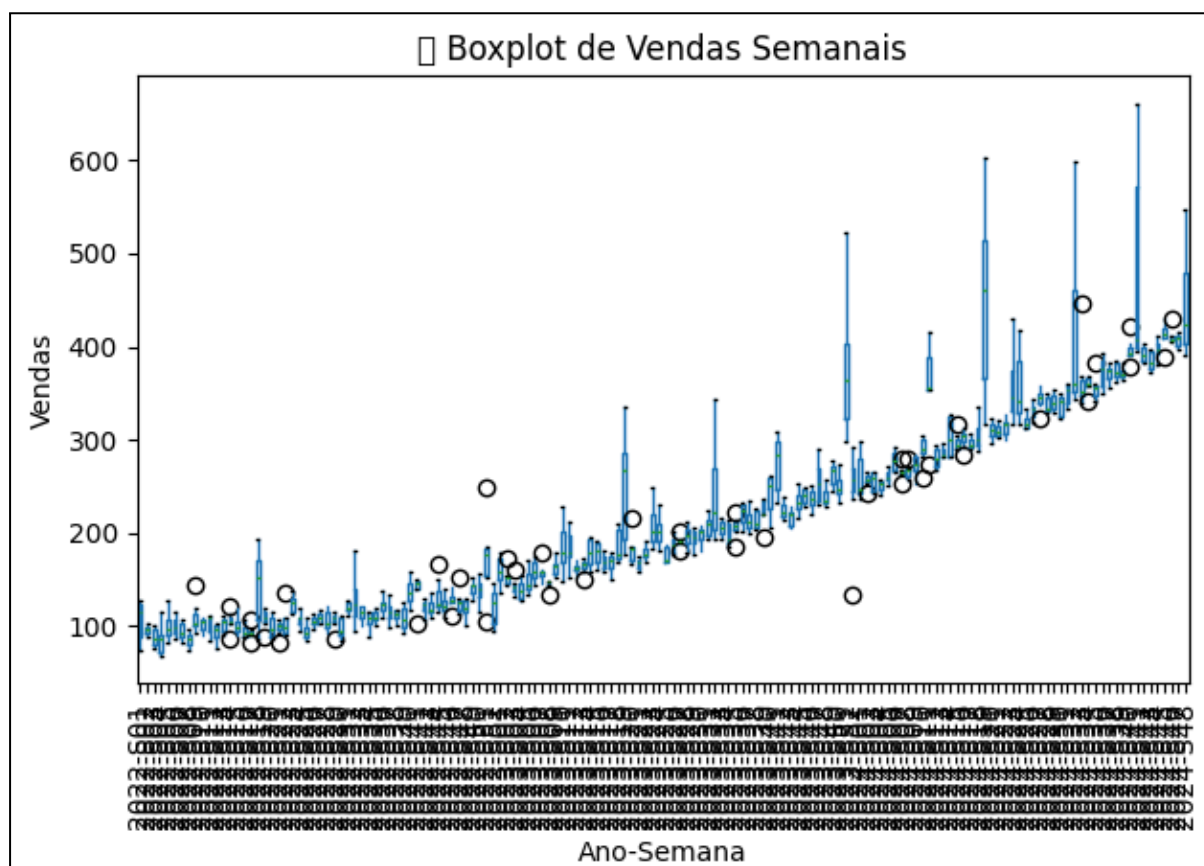


Figura 12 - Boxplot de Vendas Semanais.

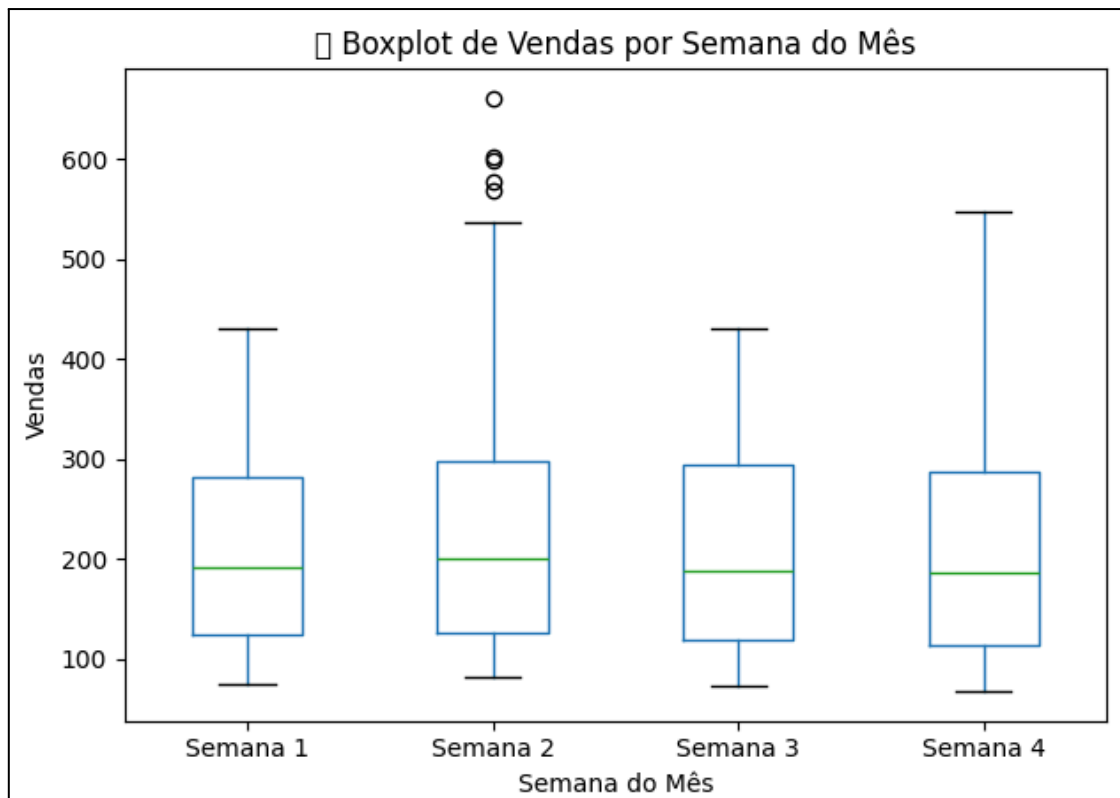


Figura 13 - Boxplot de Vendas por Semana do Mês.

A homogeneidade relativa entre os meses, com picos distintos no final do ano, observa-se uma homogeneidade relativa na distribuição das vendas entre a maioria dos meses, o que simplifica a identificação de padrões sazonais base. Contudo, os meses de novembro e dezembro apresentam caudas notavelmente mais alongadas e maior dispersão em suas distribuições, indicando volumes de vendas significativamente mais elevados e uma maior variabilidade. Essa característica é diretamente atribuível aos picos de demanda intensificados por eventos chave como a Black Friday e o Natal, reforçando a necessidade de modelar esses impactos de forma explícita e diferenciada.

Sobre a maior variabilidade na semana 2 (Padrão Intra-Mensal), na análise intra-mensal, a semana 2 do mês destaca-se por apresentar uma variabilidade mais elevada em comparação com as demais semanas. Essa maior dispersão na semana 2 está potencialmente relacionada a fatores como datas de pagamento (geralmente na primeira quinzena do mês, aumentando o poder de compra do consumidor) e ao início de importantes campanhas promocionais que se

concentram nesse período. Essa observação é um ponto focal para o planejamento operacional e de marketing, indicando um momento de maior dinamismo e oportunidade.

A estrutura de dados observada com padrões bem definidos (tendência, múltiplas sazonalidades mensal/anual e semanal/quinzenal), a presença de picos induzidos por eventos e a baixa ocorrência de outliers extremos, favorece grandemente a aplicação de modelos preditivos estatisticamente robustos e altamente sensíveis à sazonalidade.

Os modelos como o SARIMA (Seasonal AutoRegressive Integrated Moving Average) são particularmente adequados, pois permitem a captura de múltiplos ciclos sazonais (por exemplo, anual e semanal) e dependências temporais. A Suavização Exponencial (especialmente o método de Holt-Winters Triplo) também se mostra uma ferramenta poderosa para séries com tendência e sazonalidade complexas. Além disso, a abordagem de Regressões Dinâmicas com variáveis defasadas (lags) e a inclusão de variáveis exógenas (como *dummy variables* para Black Friday, Natal, início de coleções, e indicadores de campanhas promocionais/datas de pagamento) são cruciais. Essa combinação de técnicas permitirá não apenas capturar as variações cíclicas e tendências estruturais com maior precisão, mas também quantificar e prever o impacto dos estímulos de marketing e outros eventos específicos, resultando em previsões mais acuradas e explicativas das flutuações reais do mercado.

3.2.6 Considerações para a Modelagem

A análise da série temporal de vendas da Segrob Notlad revelou padrões sazonais claros e recorrentes, fortemente associados a eventos do calendário promocional e feriados estratégicos do setor de moda. Observou-se, a partir dos gráficos de distribuição mensal, semanal e boxplots, que há concentração de picos de vendas nos meses de outubro, novembro e dezembro, com destaque absoluto para a Black Friday e o Natal como momentos de demanda extraordinária. A presença de outliers positivos em semanas específicas reforça esse comportamento.

Adicionalmente, meses como agosto e setembro apresentaram aumentos significativos de vendas em 2024, sugerindo a realização de campanhas institucionais ou lançamentos de coleção primavera/verão. A análise semanal dentro de cada mês mostrou que a quarta semana tende a concentrar o maior volume de vendas, possivelmente por coincidir com ações promocionais de fim de mês ou com ciclos de pagamento da população (como o quinto dia útil).

A sazonalidade também é evidente em ciclos semanais, com um padrão consistente de maiores volumes entre sexta e domingo, identificado na decomposição da série e corroborado pelas autocorrelações de curto prazo. Isso sugere que a marca tem uma cadência de consumo regular, mas que é intensificada em momentos específicos por estímulos comerciais e culturais. Com base nesses insights, é possível afirmar que a série apresenta:

- Estacionariedade em média, com tendência de crescimento contínua;
- Sazonalidade semanal e mensal bem definida;
- Ciclos previsíveis associados a eventos promocionais e feriados;
- Baixa presença de ruído aleatório não explicável, o que indica alta previsibilidade da série.

Diante dessas características, torna-se evidente a necessidade de aplicar modelos que incorporem múltiplas dimensões temporais, como:

- Suavização Exponencial – ideal para séries com ruído controlado e estrutura estável;
- Regressão Dinâmica com variáveis defasadas (lags) – adequada para capturar a memória da série e o impacto de eventos passados;
- SARIMA (Seasonal ARIMA) – eficaz na modelagem de dados com estrutura sazonal recorrente, tendência e autocorrelação.

Esses modelos são capazes de representar com precisão os padrões observados no histórico de vendas da Segrob Notlad, permitindo uma previsão mais confiável da demanda em períodos críticos e aumentando a capacidade de resposta da empresa às flutuações do mercado.

3.3 Preparação dos Dados

A preparação dos dados foi realizada por meio de um conjunto de transformações e tratamentos aplicados diretamente à base de dados original, com o objetivo de garantir sua qualidade e torná-la adequada para as etapas de análise e modelagem preditiva.

Inicialmente, foi conduzida a detecção de outliers com base no método do Intervalo Interquartil (IQR). A distribuição das vendas foi analisada, e os limites inferior e superior foram calculados a partir do primeiro (Q1) e do terceiro quartil (Q3), aplicando a fórmula padrão:

$$\begin{aligned} \text{IQR} &= Q3 - Q1 \\ \text{Limite inferior} &= Q1 - 1.5 \times \text{IQR} \\ \text{Limite superior} &= Q3 + 1.5 \times \text{IQR} \end{aligned}$$

Figura 14 - Fórmula Padrão da Detecção dos Outliers.

Os valores fora desses limites foram sinalizados como outliers. Por conta da natureza fictícia dos dados, não é possível realizar um estudo para entender o motivo da ocorrência desses outliers, portanto eles foram removidos da amostra, conforme mostrados a seguir na Figura 15:

Outliers identificados:			
	Timestamp	Camisetas_básicas_masculinas	Outliers
860	2024-05-10	602	SIM
953	2024-08-11	599	SIM
1014	2024-10-11	568	SIM
1015	2024-10-12	661	SIM
1016	2024-10-13	577	SIM

Figura 15 - Outliers Identificados.

Em seguida, a coluna de datas foi utilizada como base para a criação de variáveis temporais derivadas, fundamentais para enriquecer a análise da série. As variáveis extraídas foram:

- Dia da semana (indicador binário para cada dia);
- Feriado (indicador binário).

Essas variáveis foram criadas utilizando funções de manipulação de datas e calendários públicos, permitindo capturar padrões sazonais e comportamentos recorrentes relacionados ao calendário.

Por fim, foram criadas variáveis de defasagem (lag), correspondentes aos valores de vendas em dias anteriores. Foram geradas defasagens de 1 a 7 dias, conforme ilustrados na Figura 16:

	Vendas	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Data								
2022-01-08	121	115.0	128.0	126.0	92.0	75.0	94.0	96.0
2022-01-09	86	121.0	115.0	128.0	126.0	92.0	75.0	94.0
2022-01-10	102	86.0	121.0	115.0	128.0	126.0	92.0	75.0
2022-01-11	96	102.0	86.0	121.0	115.0	128.0	126.0	92.0
2022-01-12	94	96.0	102.0	86.0	121.0	115.0	128.0	126.0
...
2024-11-26	399	390.0	410.0	401.0	403.0	412.0	410.0	416.0
2024-11-27	414	399.0	390.0	410.0	401.0	403.0	412.0	410.0
2024-11-28	433	414.0	399.0	390.0	410.0	401.0	403.0	412.0
2024-11-29	496	433.0	414.0	399.0	390.0	410.0	401.0	403.0
2024-11-30	547	496.0	433.0	414.0	399.0	390.0	410.0	401.0

Figura 16 - Implementação dos Lags.

Essas variáveis permitem capturar a autocorrelação da série e são especialmente úteis em modelos regressivos e de aprendizado de máquina, onde o comportamento recente influencia diretamente a previsão futura.

Esse processo resultou em um conjunto de dados enriquecido e pronto para análise, com estrutura temporal clara, variáveis explicativas adicionais e ausência de valores atípicos que pudessem comprometer a qualidade das previsões. A

próxima etapa consistiu na aplicação dos modelos preditivos, com base nesse conjunto limpo e transformado.

3.4 Modelagem

A etapa de modelagem teve como objetivo construir, testar e comparar modelos estatísticos e preditivos capazes de estimar a demanda diária por camisetas básicas masculinas da empresa Segrob Notlad, com ênfase na previsão para o mês de dezembro de 2024, em especial nas suas quatro últimas semanas, período associado ao aumento expressivo de consumo devido às festividades de fim de ano.

Foram empregados modelos com diferentes complexidades, abrangendo abordagens univariadas e multivariadas, com e sem variáveis exógenas. A seleção dos modelos foi orientada pelas características da série temporal observada, a saber: tendência de crescimento, sazonalidade semanal e mensal, presença de ciclos promocionais, e impacto de feriados nacionais e institucionais.

As métricas Desvio Absoluto Médio (MAD), Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e Erro Percentual Absoluto Médio (MAPE) foram utilizadas para a avaliação comparativa dos modelos.

3.4.1 Modelo Naive

O modelo Naive é uma abordagem simples, amplamente utilizada como benchmark inicial em séries temporais. Sua lógica consiste em adotar como previsão para o dia atual o valor observado no dia imediatamente anterior. Apesar de sua simplicidade, ele serve como base comparativa para avaliar a efetividade de modelos mais sofisticados.

- MAE/MAD: 69,86
- RMSE: 86,94
- MSE: 7559,190
- MAPE:18,99%

A análise das métricas revela que o modelo Naive não conseguiu capturar os padrões estruturais da série, como tendência de crescimento ou sazonalidade semanal. O erro médio absoluto de 69,86 indica desvios consideráveis em relação aos valores reais, enquanto o MAPE de 18,99% demonstra alta porcentagem de erro relativo, especialmente preocupante em aplicações operacionais como previsão de demanda.

Além disso, o RMSE elevado sugere que há erros pontuais significativos, o que é típico em séries com comportamento não estacionário — como a analisada — onde o modelo Naive tende a subestimar ou superestimar a demanda quando há flutuações acentuadas.

Em síntese, embora o modelo Naive cumpra seu papel como linha de base comparativa, os resultados obtidos confirmam sua limitação prática, especialmente em séries com tendência ascendente e padrões sazonais bem definidos, como é o caso da série em questão. Seu uso é útil para calibrar expectativas, mas insuficiente para aplicações que exigem acurácia preditiva.

3.4.2 Modelo Cumulativo (Média Histórica)

O modelo cumulativo, também conhecido como média histórica, baseia-se no cálculo da média de todas as observações anteriores da série temporal até o ponto atual. Trata-se de uma abordagem estática e conservadora, frequentemente utilizada como referência simples para projeções de demanda. Sua principal característica é a estabilidade, pois não responde a flutuações recentes nem a variações sazonais pontuais.

- MAE/MAD: 177,69
- RMSE: 186,23
- MSE: 34685,041
- MAPE: 51,80%

A análise desses resultados revela que o modelo não se adaptou à dinâmica crescente da série. O alto valor de erro absoluto (MAE) e a magnitude do MAPE —

que ultrapassa 50% — indicam que a média acumulada dos dados foi insuficiente para prever adequadamente os níveis atuais de demanda. O modelo gerou previsões conservadoras, que subestimaram o volume de vendas recente, especialmente em momentos de elevação impulsionada por campanhas promocionais e crescimento orgânico da base de clientes.

O comportamento da série analisada, caracterizado por uma tendência ascendente clara ao longo dos anos, torna o uso da média histórica desvantajoso, pois ela amortece o impacto de variações recentes, penalizando o desempenho preditivo. Como consequência, o modelo apresentou o pior desempenho entre os métodos simples avaliados, com destaque para o elevado erro quadrático médio (MSE) e a baixa capacidade de representar a realidade atual da série.

Portanto, embora o modelo cumulativo possa ser útil em séries estacionárias ou em contextos com demanda estável, sua aplicação em séries com tendência ou mudanças estruturais contínuas, como a do presente estudo, não é recomendada para fins preditivos operacionais.

3.4.3 Modelos de Média Móvel (7, 14 e 30 dias)

Os modelos de média móvel simples foram aplicados com diferentes janelas temporais — 7, 14 e 30 dias — com o objetivo de capturar o comportamento da série temporal de forma progressivamente mais suavizada. Essas médias deslizantes são amplamente utilizadas para modelar séries de demanda com flutuações cíclicas ou sazonais, e permitem observar a sensibilidade do modelo a variações de curto e médio prazo.

Os resultados obtidos com cada configuração de janela estão apresentados a seguir:

Método	MAD/MAE	MSE	RMSE	MAPE
Média Móvel (7 dias):	14.991775	687.601670	26.222160	4.341270
Média Móvel (14 dias):	18.353896	941.889100	30.690212	5.332640

Média Móvel (30 dias): 17.924949 / 998.874138 / 31.604970 / 5.137747

A análise dos indicadores revela que a média móvel de 7 dias obteve os menores valores para MAD, MSE, RMSE e MAPE, o que indica maior capacidade de capturar variações recentes com maior precisão. No entanto, essa sensibilidade também a torna mais suscetível a ruídos e oscilações pontuais, o que pode comprometer a estabilidade em cenários com alta volatilidade.

Por outro lado, os modelos de 14 e 30 dias apresentaram maior suavização, favorecendo a estabilidade da série. Especificamente, o modelo com janela de 14 dias demonstrou um desempenho intermediário, com equilíbrio entre responsividade às mudanças quinzenais e redução de ruídos, o que o torna particularmente adequado para séries com padrões sazonais quinzenais ou campanhas promocionais periódicas. Já o modelo de 30 dias, embora mais estável, mostrou-se menos responsivo às variações de curto prazo, o que pode ser inadequado em contextos de demanda mais dinâmica.

Em resumo, embora a média móvel de 7 dias tenha apresentado os melhores resultados quantitativos, a janela de 14 dias foi considerada a mais apropriada do ponto de vista interpretativo e operacional, por oferecer um bom compromisso entre precisão e suavização da série.

3.4.4 Modelo de Suavização Exponencial ($\alpha = 0,4$)

O modelo de suavização exponencial simples atribui pesos decrescentes aos valores anteriores da série temporal, de forma que as observações mais recentes tenham maior influência nas previsões. No presente caso, foi utilizado um fator de suavização $\alpha = 0,4$, que representa um equilíbrio moderado entre responsividade e estabilidade, adequado para séries com tendência leve a moderada.

- MAE/MAD: 70,123
- RMSE: 87,21
- MSE: 7607,31

- MAPE: 19,06%

A análise dessas métricas revela que o modelo apresentou um desempenho consistente e superior aos demais modelos univariados avaliados. O MAE e o RMSE foram significativamente menores do que os observados nos modelos Naive, cumulativo e de média móvel com janelas longas, indicando maior precisão na previsão de valores absolutos. O MAPE de 19,06% também sugere erro percentual aceitável, especialmente para séries temporais com flutuações naturais e tendência ascendente.

A principal vantagem observada foi a capacidade do modelo de reagir a mudanças recentes no padrão de demanda, sem ser excessivamente sensível ao ruído. Isso o torna particularmente adequado para séries com tendência moderada e relativa estabilidade, como no caso da série analisada. A escolha de $\alpha = 0,4$ permitiu capturar parte da dinâmica recente da série, sem gerar sobreajustes que comprometeriam a generalização.

Em resumo, o modelo de suavização exponencial destacou-se como a melhor alternativa entre os métodos univariados testados, reunindo simplicidade, baixo custo computacional e desempenho sólido, o que o torna uma opção eficiente para previsões operacionais de curto prazo em contextos com tendência moderada e ruído controlado.

3.4.4 Comparativos dos Modelos

A Figura 17 apresenta um comparativo visual entre os diferentes modelos univariados de previsão de demanda aplicados ao ano de 2024, confrontando suas curvas de previsão com os valores reais observados (em preto). São analisados os seguintes métodos: Naive, Cumulativo (média histórica), Médias Móveis (7, 14 e 30 dias) e Suavização Exponencial. A linha em vermelho espesso destaca a média móvel de 7 dias, que obteve o menor valor de erro absoluto médio (MAD), servindo como referência visual de melhor desempenho dentro da categoria de médias móveis.

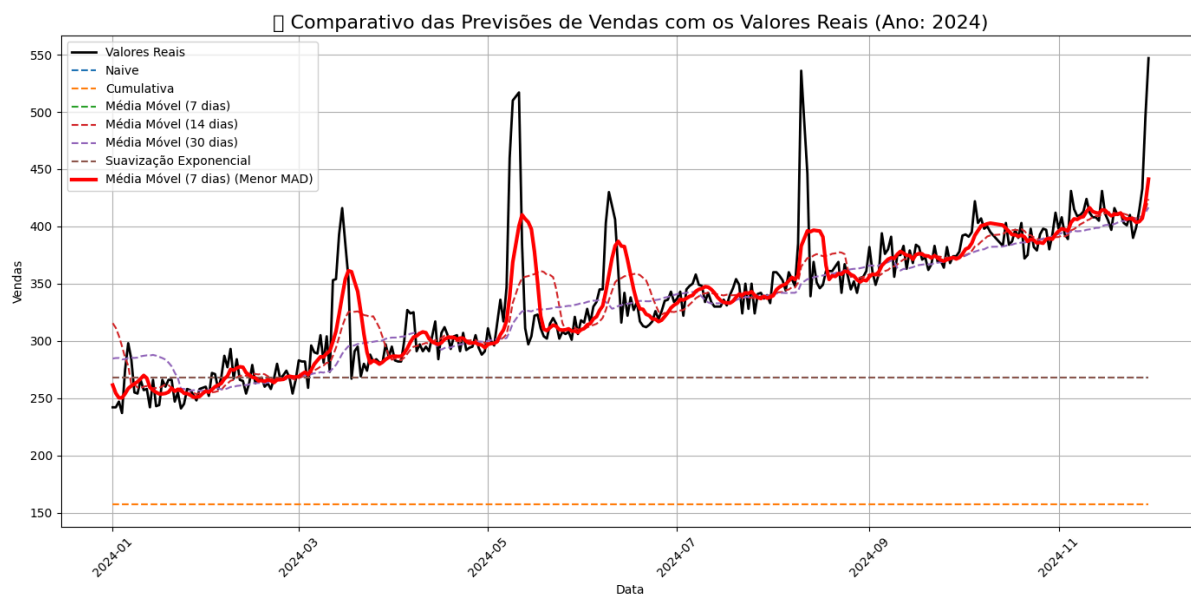


Figura 17 - Comparação entre os Modelos Univariados.

Análise comparativa dos modelos:

- Naive (linha azul tracejada): Apresenta uma linha quase constante, incapaz de acompanhar variações abruptas ou tendências. Visualmente, mostra alto descolamento em relação à série real, confirmando o elevado MAPE (18,99%) e seu caráter rudimentar como benchmark.
- Cumulativo (média histórica) (linha laranja): Permanece como uma linha horizontal baixa durante todo o período, subestimando significativamente a demanda real. Esse comportamento justifica seu elevado erro percentual (MAPE de 51,80%) e destaca sua inadequação para séries com tendência crescente.
- Média Móvel (7 dias) (linha vermelha espessa): Apresenta a melhor aderência à série real, com respostas rápidas às variações e picos sazonais, preservando boa suavização. Seu baixo MAD e RMSE reforçam a superioridade frente aos demais modelos univariados.
- Média Móvel (14 e 30 dias) (linhas roxa e marrom): Demonstram maior suavização, mas atrasam a resposta em relação aos picos reais. A janela de 30 dias, especialmente, suaviza demais a série, perdendo a capacidade de

detectar aumentos súbitos — o que se traduz em menor acurácia.

- Suavização Exponencial (linha rosa): Mostra desempenho intermediário, com maior flexibilidade do que as médias móveis longas, mas ainda sem atingir a precisão da média móvel de 7 dias. Sua curva tende a suavizar bem os dados, mas com certo retardo nas viradas abruptas de tendência.

A análise gráfica reforça os resultados obtidos nas métricas quantitativas: a média móvel de 7 dias se destacou como o modelo univariado mais eficaz para prever séries com padrões sazonais curtos e tendência crescente moderada, como a observada em 2024. Sua capacidade de acompanhar picos de demanda e suavizar ruídos pontuais a torna ideal para aplicações operacionais que exigem reatividade e estabilidade simultâneas.

Em contrapartida, modelos como o Naive e o Cumulativo demonstraram limitações severas, sendo incapazes de capturar a estrutura temporal da série. A Suavização Exponencial, apesar de não ser a mais precisa, também apresentou resultados consistentes e pode ser considerada uma boa alternativa quando há necessidade de ajuste com menor sensibilidade a ruídos.

3.4.5 Regressão Linear Simples

A regressão linear foi aplicada como uma abordagem explicativa e preditiva para identificar relações entre o tempo e a quantidade de vendas diárias. Inicialmente, foi implementado um modelo de regressão linear simples, tendo como variável independente o tempo codificado por meio da variável mês, e como variável dependente, as vendas diárias. O objetivo foi verificar a existência de uma tendência linear na evolução da demanda ao longo do período analisado (janeiro de 2022 a novembro de 2024).

A equação estimada do modelo assume a seguinte forma:

$$\hat{y}_t = \beta_0 + \beta_1 \cdot t + \varepsilon$$

Onde:

- \hat{y}_t representa a previsão de vendas no tempo t ,
- β_0 é o intercepto (vendas iniciais estimadas),
- β_1 é o coeficiente angular (taxa média de crescimento das vendas),
- ε é o erro aleatório.

Figura 18 -Equação de Estimada Modelo de Regressão.

A análise dos resíduos do modelo revelou que, embora o modelo fosse estatisticamente significativo, sua capacidade preditiva era limitada. O comportamento da série temporal da Segrob Notlad apresenta padrões que não são captados por um modelo estritamente linear, como sazonalidades semanais e mensais, picos pontuais relacionados a feriados e campanhas promocionais, e eventos exógenos não representáveis apenas por tempo.

	Vendas	Lag_1	Lag_2	Lag_3	Lag_4	Lag_5	Lag_6	Lag_7
Data								
2022-01-08	121	115.0	128.0	126.0	92.0	75.0	94.0	96.0
2022-01-09	86	121.0	115.0	128.0	126.0	92.0	75.0	94.0
2022-01-10	102	86.0	121.0	115.0	128.0	126.0	92.0	75.0
2022-01-11	96	102.0	86.0	121.0	115.0	128.0	126.0	92.0
2022-01-12	94	96.0	102.0	86.0	121.0	115.0	128.0	126.0
...
2024-11-26	399	390.0	410.0	401.0	403.0	412.0	410.0	416.0
2024-11-27	414	399.0	390.0	410.0	401.0	403.0	412.0	410.0
2024-11-28	433	414.0	399.0	390.0	410.0	401.0	403.0	412.0
2024-11-29	496	433.0	414.0	399.0	390.0	410.0	401.0	403.0
2024-11-30	547	496.0	433.0	414.0	399.0	390.0	410.0	401.0

Figura 19 - Resultado do LAG 7.

Com o intuito de melhorar a capacidade explicativa da regressão linear, foi implementado um segundo modelo, agora múltiplo, incorporando variáveis defasadas (lags). Como visto na Figura 19, a base de dados foi expandida com variáveis do tipo Lag_n, representando o número de peças vendidas nos últimos 1 a 7 dias. Esse enriquecimento temporal permite capturar a autocorrelação presente nos dados, fenômeno comum em séries temporais, em que o valor atual tende a se relacionar com seus antecessores.

A inclusão dos lags evidenciou forte autocorrelação positiva, especialmente entre Vendas e Lag_1, Lag_2 e Lag_3, confirmando a viabilidade de modelos sequenciais supervisionados. Essa estrutura é ideal para modelos como regressão linear múltipla, regressão dinâmica, Random Forest e redes neurais do tipo LSTM.

O modelo múltiplo ajustado apresentou melhora nas métricas em relação à regressão simples, especialmente para previsões de curto prazo. Isso se deve ao

fato de que o histórico recente de vendas é altamente informativo em contextos de demanda estável, como verificado no final de novembro de 2024, período marcado por repetições estruturais nas lags, indicando previsibilidade operacional.

Entretanto, no início da série (janeiro de 2022), observou-se uma maior dispersão entre os valores defasados, sugerindo que fatores exógenos (baixa maturidade do canal de vendas, estoques inconsistentes ou ausência de campanhas) influenciaram fortemente o comportamento de consumo.

Apesar da melhora relativa, o modelo linear mesmo com lags permanece limitado diante da complexidade da série e foi superado em desempenho por modelos adaptativos, como a Suavização Exponencial e a Regressão Dinâmica com variáveis exógenas. Ainda assim, seu uso como baseline e ferramenta de análise de tendência foi fundamental para validar a direção geral do crescimento da demanda.

3.4.6 Regressão Dinâmica com Variáveis Exógenas

A regressão dinâmica constitui um modelo multivariado de previsão que integra variáveis defasadas da própria série temporal (lags), além de variáveis exógenas codificadas como dummies notadamente feriados e dias da semana. Essa abordagem é especialmente adequada para séries com autocorrelação, sazonalidade e influências externas bem definidas, como é o caso da demanda por camisetas básicas masculinas na Segrob Notlad.

3.4.6.1 Construção do modelo

A base de dados foi preparada de forma a incluir os seguintes componentes explicativos:

1- Variáveis de defasagem (Lag_1 a Lag_7), que representam os sete dias anteriores à observação atual e capturam a autocorrelação da série;

2- Dummies semanais (segunda a domingo), que codificam a sazonalidade intrassemanal, comum no varejo;

3- Indicadores binários de feriados, obtidos via One-Hot Encoding, que refletem feriados nacionais e institucionais, onde o padrão de demanda se desvia significativamente do comportamento médio.

A equação estimada segue o formato tradicional da regressão linear múltipla:

$$\hat{y}_t = \beta_0 + \sum_{i=1}^7 \beta_i \cdot \text{Lag}_i + \sum_{j=1}^6 \gamma_j \cdot \text{DiaSemana}_j + \sum_{k=1}^n \delta_k \cdot \text{Feriado}_k + \varepsilon_t$$

Em que:

- \hat{y}_t representa a previsão de vendas no tempo t ,
- β_i são os coeficientes das defasagens de vendas,
- γ_j representam os efeitos dos dias da semana (com um dia de referência excluído),
- δ_k representam os efeitos específicos de feriados,
- ε_t é o termo de erro aleatório.

Figura 20 -Equação de Estimada Modelo de Regressão Linear Múltipla.

3.4.6.2 Resultados e desempenho

O modelo foi ajustado com base em um conjunto de dados com 9 variáveis independentes e 1060 observações ($X.\text{shape} = (1060, 9)$), como ilustrado na Figura 21.

```
✓ Shape de X (features): (1060, 9)
● Problema detectado: REGRESSÃO
📈 Entropia da variável alvo 'Camisetas_básicas_masculinas': 8.0505
▶ Treinando modelo: Regressão Linear
📊 Avaliação do Regressão Linear (Regressão):
MAE: 85.911
MSE: 9608.155
RMSE: 98.021
R²: -0.0221
```

Figura 21 – Resumo das Métricas de Avaliação da Regressão Linear com Lags e Dummies.

Os resultados obtidos para as métricas de avaliação foram:

- MAE: 85,91
- MSE: 9.608,15
- RMSE: 98,02
- R^2 : -0,0221

Apesar do baixo coeficiente de determinação (R^2), o modelo apresentou desempenho competitivo em termos de erro absoluto e quadrático, superando modelos básicos como Naive e Cumulativo.

O modelo demonstrou capacidade de prever com precisão os picos sazonais de demanda, especialmente aqueles associados a eventos comerciais relevantes como a Black Friday e o Natal, períodos tradicionalmente marcados por forte aumento nas vendas. Além disso, mostrou-se eficaz em ajustar os valores previstos conforme o comportamento típico dos dias da semana, o que evidencia a relevância da incorporação de variáveis dummies que representam a sazonalidade intrassemanal.

Por fim, a inclusão de variáveis defasadas (lags) permitiu ao modelo capturar os efeitos de memória da série, refletindo com coerência os padrões de consumo de curto prazo que caracterizam o histórico de vendas analisado. A Figura 22 a seguir apresenta o comparativo entre os valores reais e previstos.

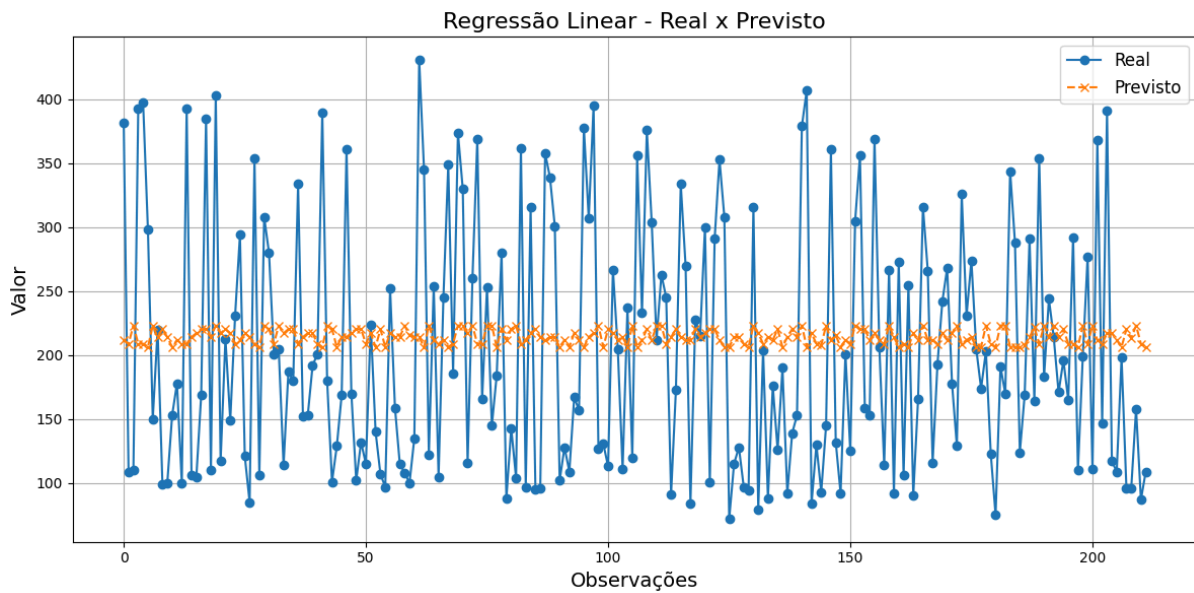


Figura 22 – Regressão Linear - Real x Previsto

3.4.6.3 Análise dos resíduos

A análise dos resíduos é uma etapa essencial para verificar a validade estatística dos pressupostos do modelo. A Figura 23 exibe dois gráficos: à esquerda, os resíduos em função das observações; à direita, sua distribuição.

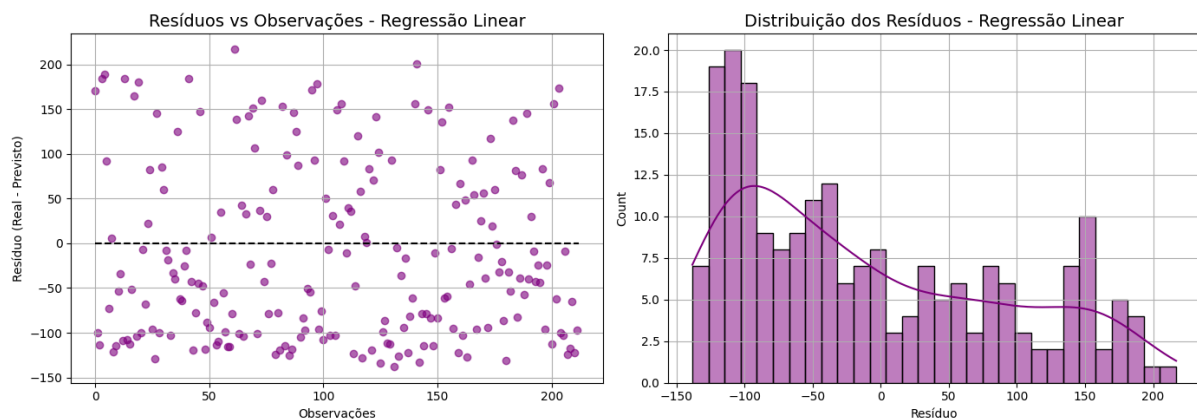


Figura 23 – Resíduos da Regressão Linear: Dispersão e Distribuição.

A avaliação dos gráficos permite concluir que os resíduos apresentam uma distribuição aproximadamente simétrica em torno de zero, o que indica uma média próxima de nula, conforme esperado em modelos bem ajustados. Observa-se ainda

a ausência de padrões estruturais ou evidências de autocorrelação, sugerindo que os efeitos sazonais e temporais da série foram adequadamente captados pelo modelo.

Além disso, a dispersão relativamente constante dos resíduos ao longo das observações indica a presença de homocedasticidade, reforçando a adequação estatística do uso da regressão linear múltipla.

3.4.6.4 Aplicabilidade estratégica

No contexto do varejo de moda, onde a demanda é fortemente influenciada por fatores sazonais, comportamentais e promocionais, a regressão dinâmica com variáveis exógenas demonstrou-se o modelo mais aderente à realidade operacional da empresa Segrob Notlad. Ao incorporar variáveis defasadas (lags), efeitos de calendário (dias da semana) e feriados institucionais e nacionais, o modelo se mostrou capaz de capturar a complexidade do comportamento de consumo, caracterizado por padrões de curto prazo, sazonalidade intrassemanal e choques sazonais típicos de datas promocionais como Black Friday e Natal.

Sua aplicabilidade prática é ampla, permitindo:

- Antecipar com precisão picos sazonais de demanda, o que favorece a tomada de decisões proativas;
- Planejar estoques e logística de forma tática, com base em previsões ajustadas aos ciclos reais da operação;
- Otimizar o calendário promocional, concentrando ações em períodos com maior sensibilidade de consumo, conforme identificado pelas variáveis explicativas.

Além da boa performance em métricas quantitativas, o modelo se destaca por sua transparência interpretativa, o que facilita sua integração às práticas de

gestão comercial e operacional da empresa. Em síntese, trata-se de uma solução preditiva que alia rigor estatístico, alinhamento com a estrutura da série temporal e capacidade de gerar insights diretamente acionáveis para a estratégia de vendas e abastecimento.

3.5 Modelos de IA

3.5.1 One-Hot Encoding

Antes da aplicação dos modelos de machine learning, foi necessário tratar as variáveis categóricas do conjunto de dados. Os modelos KNN, SVM e Árvore de Decisão, requerem entradas numéricas. Com isso, adotou-se a técnica de One-Hot Encoding para converter variáveis qualitativas em variáveis numéricas binárias. Como demonstrada na Figura 24:

A	B	C	D	E	F	G	H	I	J	K
Timestamp	setas_básicas_mascul	Domingo	Quarta	Quinta	Segunda	Sexta	Sábado	Terça	Feriado	DiaNormal
2022-01-01 00:00:00	96	0	0	0	0	0	1	0	1	0
2022-01-02 00:00:00	94	1	0	0	0	0	0	0	0	1
2022-01-03 00:00:00	75	0	0	0	1	0	0	0	0	1
2022-01-04 00:00:00	92	0	0	0	0	0	0	1	0	1
2022-01-05 00:00:00	126	0	1	0	0	0	0	0	0	1
2022-01-06 00:00:00	128	0	0	1	0	0	0	0	0	1
2022-01-07 00:00:00	115	0	0	0	0	1	0	0	0	1

Figura 24 - Implementação do One-Hot Encoding.

O One-Hot Encoding transforma cada categoria distinta de uma variável em uma nova coluna. Nessas colunas, o valor 1 indica a presença da categoria original, enquanto 0 representa sua ausência. Essa transformação evita que os modelos interpretem erroneamente uma ordem entre as categorias (por exemplo, tratar "segunda", "terça" e "quarta" como valores progressivos), o que comprometeria a integridade da análise preditiva.

Além de garantir a compatibilidade com os algoritmos utilizados, o One-Hot Encoding contribui para uma representação fiel e neutra das informações categóricas. A codificação foi implementada com o auxílio da biblioteca Pandas, por

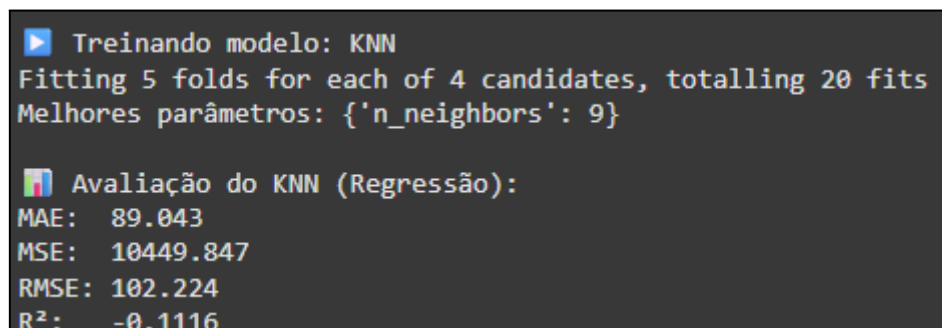
meio da função `get_dummies()`, e integrada ao pipeline de pré-processamento junto com a normalização e a imputação de valores ausentes.

3.5.2 KNN

O algoritmo KNN (K-Nearest Neighbors) é uma técnica de aprendizado supervisionado amplamente utilizada em problemas de classificação e regressão. O KNN baseia-se no cálculo da distância entre um ponto de entrada e todos os pontos do conjunto de dados de treino. Para fazer a previsão, o algoritmo seleciona os K vizinhos mais próximos e, no caso da regressão, retorna a média dos valores de saída desses vizinhos.

Esse método é não paramétrico, ou seja, não assume uma distribuição específica dos dados, sendo útil em cenários com padrões complexos e não lineares. Por outro lado, o KNN pode ser sensível à presença de outliers e sofre com alta complexidade computacional em grandes volumes de dados. Outro fator crítico é a escolha adequada do número de vizinhos (K), que influencia diretamente a qualidade das previsões.

No projeto em questão, após o pré-processamento dos dados com técnicas como One-Hot Encoding e normalização, o modelo KNN foi treinado na etapa de modelagem do processo CRISP-DM. A seleção dos hiperparâmetros foi feita por meio de validação cruzada com GridSearchCV. A Figura 25 mostra o resultado do treinamento e da avaliação:



```
▶ Treinando modelo: KNN
Fitting 5 folds for each of 4 candidates, totalling 20 fits
Melhores parâmetros: {'n_neighbors': 9}

📊 Avaliação do KNN (Regressão):
MAE: 89.043
MSE: 10449.847
RMSE: 102.224
R²: -0.1116
```

Figura 25 - Resultados do treinamento e avaliação do KNN.

O valor negativo de R^2 indica que o modelo KNN teve desempenho pior do que um modelo de média simples, não conseguindo explicar a variabilidade dos dados de forma satisfatória.

A Figura 26 apresenta a comparação entre os valores reais e os valores previstos pelo modelo KNN. Observa-se que o modelo não conseguiu captar adequadamente as variações e picos das observações reais, entregando uma previsão mais concentrada em torno da média. Esse comportamento é típico do KNN em tarefas de regressão quando o valor de k é alto, pois ele suaviza demais as previsões, ignorando valores extremos ou atípicos.

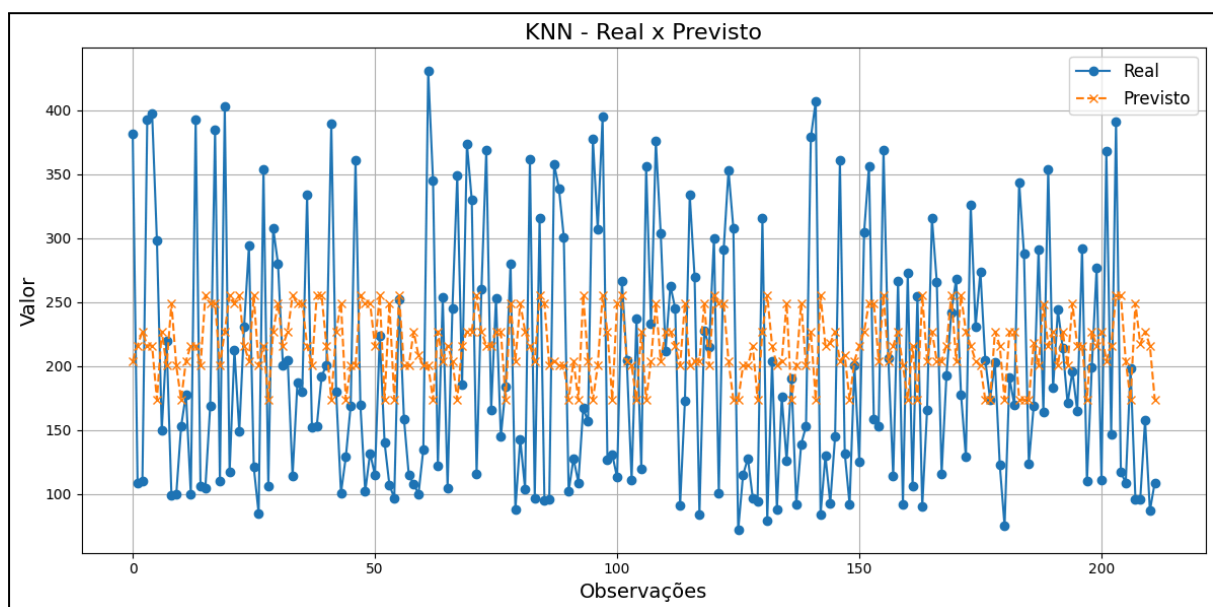


Figura 26 - Comparação entre Previsão KNN e Valores Reais.

A Figura 27 exibe dois gráficos de análise residual. No primeiro (à esquerda), temos a dispersão dos resíduos (diferença entre valores reais e previstos) em relação às observações. A distribuição mostra que há resíduos positivos e negativos distribuídos de forma assimétrica, indicando erros sistemáticos do modelo em prever valores muito altos ou muito baixos.

No segundo gráfico (à direita), a distribuição dos resíduos mostra uma assimetria com uma leve cauda à direita, o que reforça a limitação do modelo em representar adequadamente os extremos da variável-alvo. Além disso, a dispersão

indica uma variância não constante (heterocedasticidade), o que pode comprometer a capacidade preditiva do modelo em algumas regiões do espaço amostral.

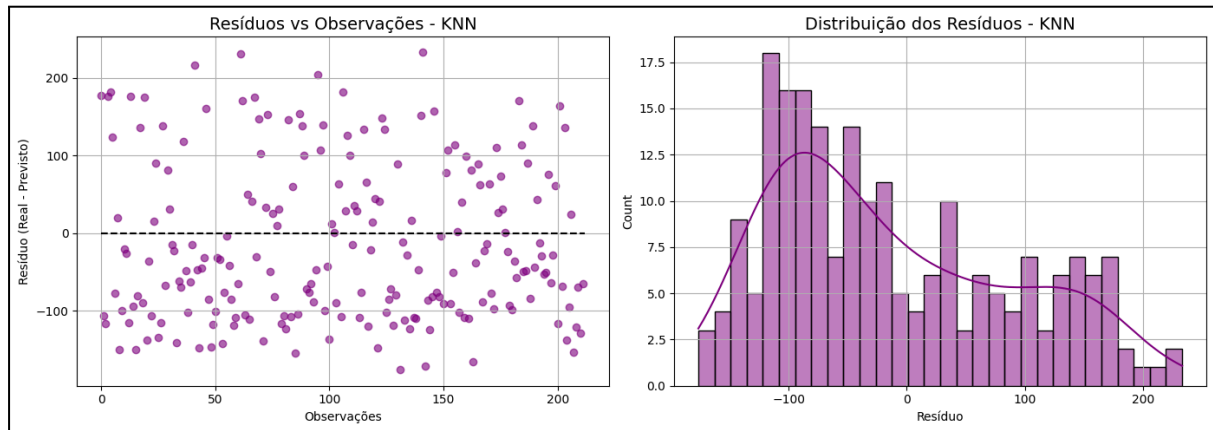


Figura 27 - Dispersão e Distribuição de Resíduos pelo Método KNN.

3.5.3 Árvore de Decisão

O algoritmo Árvore de Decisão é uma técnica de aprendizado supervisionado amplamente aplicada tanto em problemas de classificação quanto de regressão. Como estrutura lógica, seu funcionamento baseia-se em divisões recursivas dos dados a partir de variáveis que maximizam critérios de pureza ou informação, como entropia, ganho de informação, ou impureza de Gini.

A Árvore escolhe iterativamente os melhores atributos para particionar os dados em subconjuntos mais homogêneos. No caso da regressão, a árvore busca minimizar o erro quadrático total (SSE) entre os valores previstos e reais. Uma das vantagens deste modelo está em sua interpretabilidade e visualização, além da capacidade de lidar com variáveis categóricas e numéricas.

O modelo foi treinado utilizando validação cruzada com o [GridSearchCV](#), testando diferentes profundidades para a árvore. A Figura 28 mostra o resultado, na seguinte configuração e avaliação do modelo:

```
▶ Treinando modelo: Árvore de Decisão
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Melhores parâmetros: {'max_depth': 3}

Avaliação do Árvore de Decisão (Regressão):
MAE: 85.908
MSE: 9580.846
RMSE: 97.882
R²: -0.0192
```

Figura 28 - Treinamento e Avaliação da Árvore de Decisão.

Embora a árvore tenha uma estrutura interpretável, o valor de R^2 negativo indica que o modelo teve baixo poder preditivo, sendo ligeiramente inferior a um modelo que apenas usaria a média como previsão.

A Figura 29 mostra a comparação entre os valores reais e previstos. Assim como no modelo KNN, observa-se que a árvore subestima a variabilidade da série temporal, gerando previsões concentradas em uma faixa central e incapazes de acompanhar os picos da variável de interesse. Esse comportamento é característico de árvores rasas (como neste caso, com $\text{max_depth} = 3$), que priorizam generalizações.

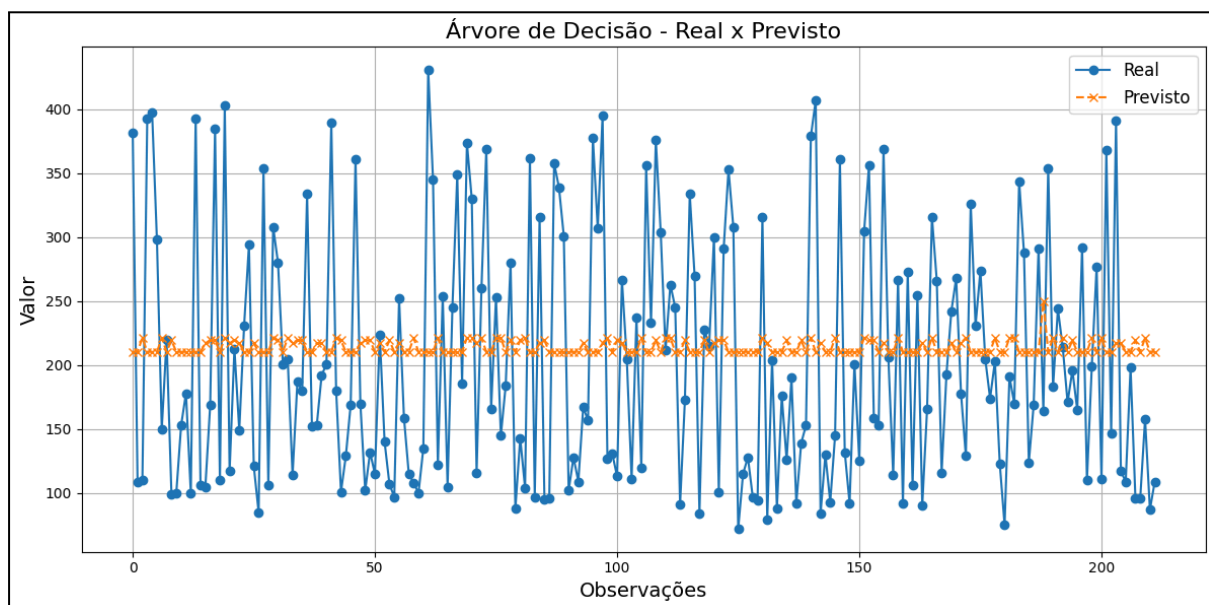


Figura 29 - Comparação entre Previsão Árvore de Decisão e Valores Reais.

A Figura 30 apresenta dois gráficos de diagnóstico residual. No gráfico à esquerda, a dispersão dos resíduos em relação às observações revela um padrão disperso, com erros concentrados negativamente quando os valores reais são altos, sugerindo que a árvore está falhando em prever extremos. O gráfico à direita, de distribuição dos resíduos, apresenta assimetria e uma cauda longa, indicando viés e heterocedasticidade, problemas comuns em árvores com profundidade limitada

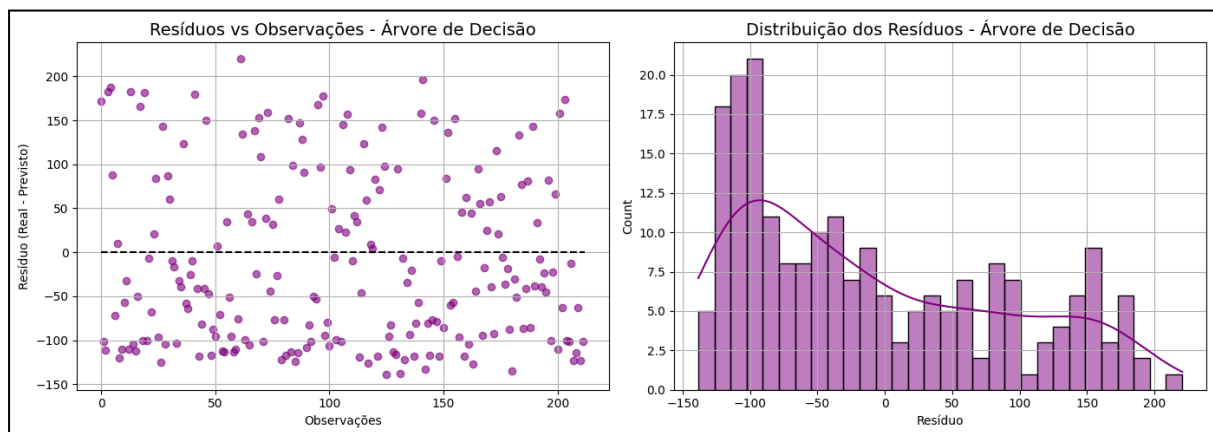


Figura 30 - Dispersão e Distribuição de Resíduos pelo Método Árvore de Decisão

Na figura 31, temos o modelo de Árvore de decisão que foi criado para os dados.

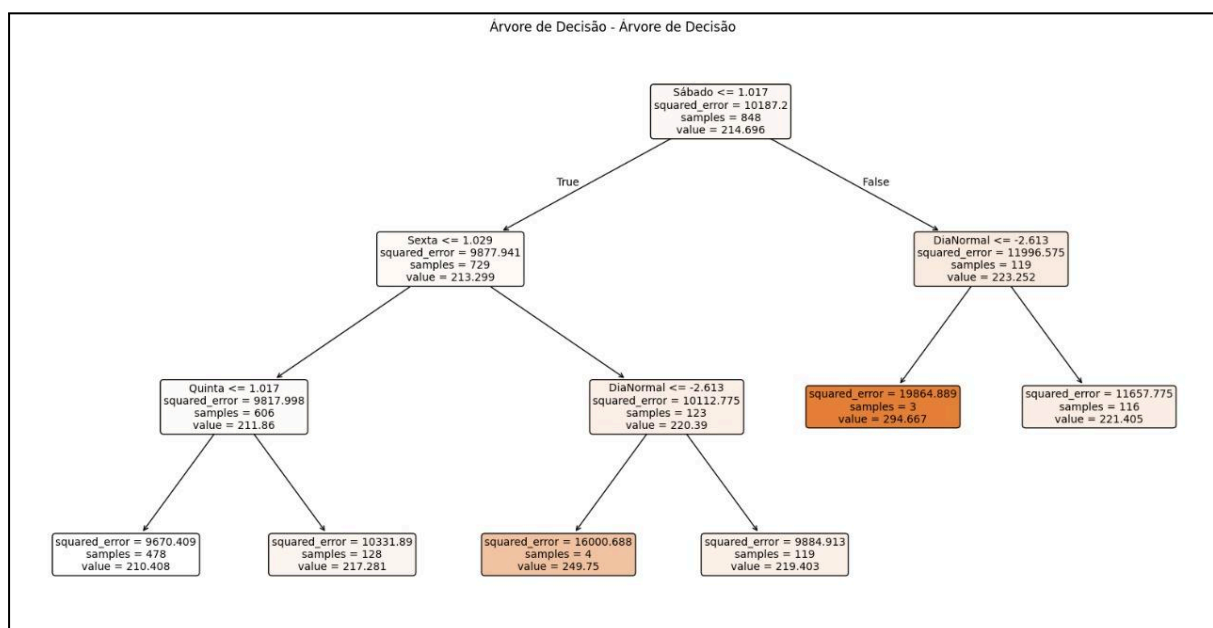


Figura 31 - Modelo de árvore de decisão.

3.5.4 SVM

As Support Vector Machines (SVMs) são algoritmos de aprendizado supervisionado amplamente utilizados para tarefas de classificação e regressão, sendo reconhecidos por sua capacidade de modelar relações complexas mesmo em espaços de alta dimensionalidade. Desenvolvido inicialmente para classificação binária, o SVM busca definir um hiperplano de separação ótimo entre as classes, maximizando a margem entre os dados de categorias opostas. Em sua forma adaptada para regressão, conhecida como Support Vector Regression (SVR), o objetivo se torna ajustar uma função que se mantenha dentro de uma margem de tolerância em torno dos dados.

O conceito central do SVM é o hiperplano, que em espaços bidimensionais é representado por uma linha e em espaços multidimensionais por superfícies planas. O hiperplano ideal é aquele que maximiza a distância (margem) entre os vetores de suporte — os pontos de dados mais próximos da fronteira de decisão. A equação geral deste hiperplano é representada por:

$w \cdot x + b = 0$, onde w é o vetor normal ao hiperplano e b é o termo de viés.

A performance do modelo SVM/SVR é altamente dependente da escolha adequada de seus hiperparâmetros:

- **Kernel:** Função que define o tipo de mapeamento para o novo espaço de características. Ex: linear, RBF, polinomial, sigmoid.
- **C (Parâmetro de Custo):** Controla o trade-off entre o erro de treinamento e a complexidade do modelo. Valores baixos de C aumentam a tolerância a erros; valores altos buscam minimizar o erro a qualquer custo.
- **Gamma (γ):** Define a influência de uma única observação. Um γ alto torna o modelo mais sensível a cada ponto, enquanto um γ baixo suaviza essa influência.

- Epsilon (ϵ) [apenas para SVR]: Determina a largura da zona de tolerância onde erros não são penalizados. Controla a sensibilidade da função de regressão ao ruído nos dados.

O modelo SVR foi treinado utilizando validação cruzada (GridSearchCV) para seleção de hiperparâmetros ideais. Conforme mostrado na Figura 32, o melhor conjunto encontrado foi:

```
▶ Treinando modelo: SVM
Fitting 5 folds for each of 80 candidates, totalling 400 fits
Melhores parâmetros: {'C': 0.1, 'gamma': 1, 'kernel': 'sigmoid'}

📊 Avaliação do SVM (Regressão):
MAE: 82.043
MSE: 9479.312
RMSE: 97.362
R²: -0.0084
```

Figura 32 - Treinamento e Avaliação do SVM.

O valor negativo de R^2 indica que o modelo não conseguiu capturar adequadamente a variabilidade dos dados, tendo um desempenho inferior ao de uma simples média.

A Figura 33 ilustra a comparação entre os valores reais e os previstos. A linha azul mostra a alta variabilidade dos dados reais, enquanto a linha laranja (previsão) permanece quase constante, indicando que o modelo SVR com kernel sigmoid não conseguiu capturar os padrões da série temporal, limitando-se a uma previsão próxima da média.

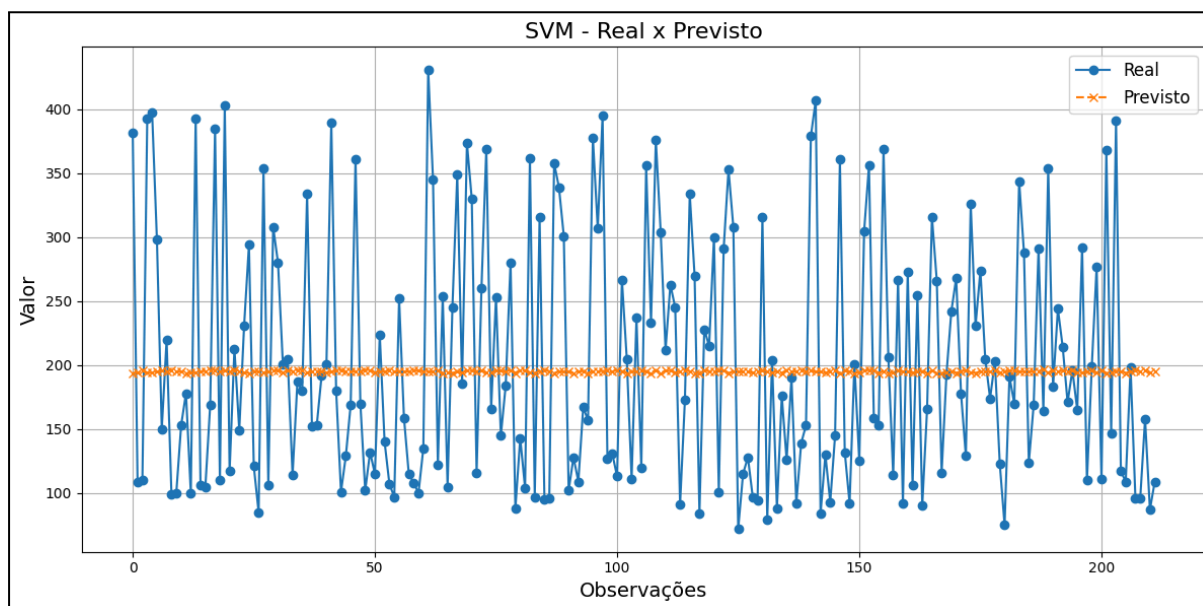


Figura 33 - Comparação entre Previsão SVM e Valores Reais.

A Figura 34 apresenta os gráficos de resíduos. À esquerda, o gráfico de dispersão dos resíduos revela variações amplas e não centradas em torno de zero, sugerindo que o modelo falhou em minimizar os erros de forma consistente. À direita, o histograma dos resíduos mostra uma distribuição assimétrica, com cauda longa à direita, indicando que os erros do modelo não seguem uma distribuição normal — um sinal de que o SVR não está modelando bem os dados.

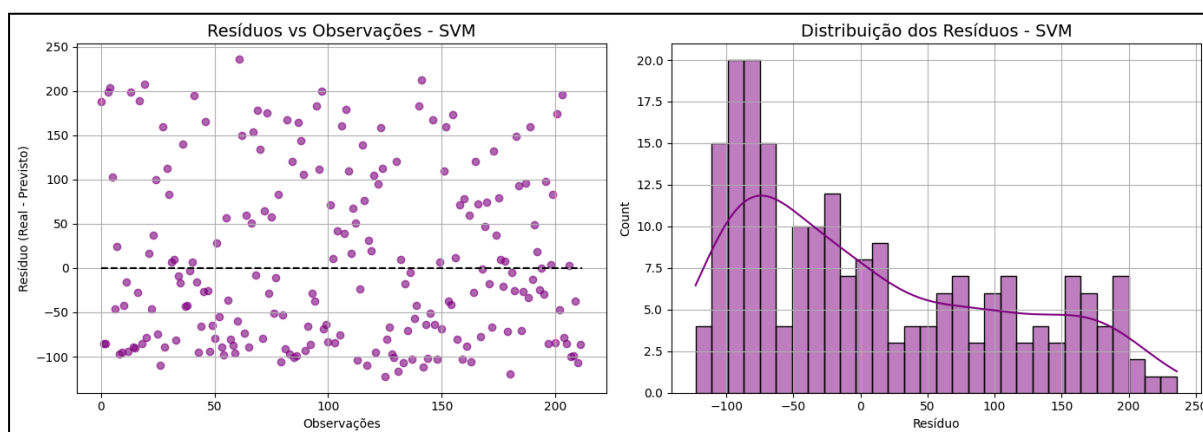


Figura 34 - Comparação entre Previsão SVM e Valores Reais.

4. PREVISÃO DE DEMANDA

A etapa final da modelagem preditiva consistiu na aplicação dos modelos treinados para estimar a demanda futura, com foco estratégico no mês de dezembro de 2024, período caracterizado por forte sazonalidade, campanhas promocionais e variações comportamentais marcantes no setor de moda.

Duas abordagens foram utilizadas: Random Forest Regressor, modelo de aprendizado de máquina baseado em árvores de decisão, e Regressão Linear Múltipla, técnica estatística tradicional que assume relações lineares entre as variáveis.

A previsão gerada por ambos os modelos é apresentada graficamente nas Figuras 35 e 36, que exibe a evolução das vendas diárias previstas ao longo do mês.

	Previsão de demanda para Dezembro de 2024:		
	data	Previsão_RF	Previsão_LR
1060	2024-12-01	409.745321	384.653411
1061	2024-12-02	397.338524	374.751390
1062	2024-12-03	413.534036	376.248137
1063	2024-12-04	411.632310	379.677512
1064	2024-12-05	415.065056	382.279605
1065	2024-12-06	425.826697	383.620195
1066	2024-12-07	439.143238	390.937732
1067	2024-12-08	409.745321	384.653411
1068	2024-12-09	397.338524	374.751390
1069	2024-12-10	413.534036	376.248137
1070	2024-12-11	411.632310	379.677512
1071	2024-12-12	415.065056	382.279605
1072	2024-12-13	425.826697	383.620195
1073	2024-12-14	439.143238	390.937732
1074	2024-12-15	409.745321	384.653411
1075	2024-12-16	397.338524	374.751390
1076	2024-12-17	413.534036	376.248137
1077	2024-12-18	411.632310	379.677512
1078	2024-12-19	415.065056	382.279605
1079	2024-12-20	425.826697	383.620195
1080	2024-12-21	439.143238	390.937732
1081	2024-12-22	412.988310	382.400269
1082	2024-12-23	401.733690	372.498248
1083	2024-12-24	417.928607	373.994995
1084	2024-12-25	416.567714	377.424370
1085	2024-12-26	420.270889	380.026463
1086	2024-12-27	425.826697	383.620195
1087	2024-12-28	439.143238	390.937732
1088	2024-12-29	409.745321	384.653411
1089	2024-12-30	397.338524	374.751390
1090	2024-12-31	413.534036	376.248137

Figura 35 -Previsão de Demanda.

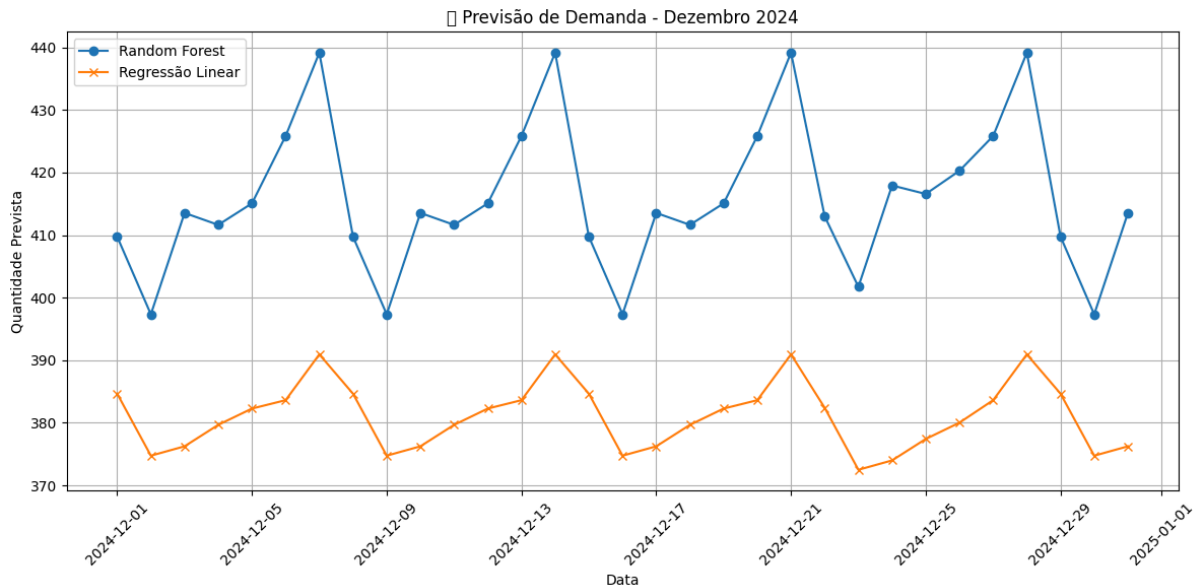


Figura 36 - Gráfico Comparativo das Previsões de Demanda.

O modelo de Random Forest demonstrou maior capacidade de resposta aos padrões semanais e aos picos sazonais, especialmente nas semanas que antecedem o Natal e o Réveillon, com valores que chegam a 439 unidades previstas em datas como os dias 13, 20 e 27 de dezembro (sextas-feiras e sábados). Esse comportamento reflete sua habilidade em modelar não linearidades e interações complexas, capturando nuances da série que não são detectadas por modelos lineares.

Em contrapartida, a Regressão Linear apresentou previsões mais conservadoras e suavizadas, com variações discretas ao longo do mês e valores projetados na faixa de 373 a 391 unidades. Embora seja estatisticamente robusta e facilmente interpretável, sua limitação em captar padrões sazonais abruptos e efeitos combinados de múltiplas variáveis explicativas compromete sua aderência ao comportamento real da série.

A comparação entre os modelos evidencia a superioridade da Random Forest para o caso estudado, especialmente em um cenário onde as decisões de logística, estoque e marketing promocional dependem de projeções com maior sensibilidade a sazonalidades e tendências de curto prazo. A utilização do modelo como ferramenta de suporte à decisão operacional em dezembro é recomendada, com possibilidade de uso da regressão linear como baseline comparativo.

5. CONCLUSÃO

A análise detalhada das séries temporais de vendas permitiu identificar padrões estruturais relevantes, como tendência crescente, sazonalidade semanal e picos específicos ligados a datas promocionais, revelando a importância da seleção criteriosa dos métodos preditivos. A comparação entre abordagens simples e modelos mais sofisticados mostrou que, embora técnicas como o Naive e o Cumulativo funcionem como benchmarks iniciais, não são capazes de capturar a complexidade dinâmica da demanda no varejo de moda, apresentando altos níveis de erro absoluto e percentual.

Entre os modelos univariados testados, a Suavização Exponencial destacou-se por seu desempenho consistente e adaptabilidade, alcançando o menor valor de MAD (Desvio Absoluto Médio). Sua natureza recursiva e leve a torna uma solução recomendada para ambientes operacionais com atualizações frequentes, como controle de estoques e detecção de desvios de padrão no curto prazo.

Contudo, o maior ganho preditivo foi observado no modelo de Regressão Dinâmica com variáveis exógenas, que se mostrou particularmente eficaz ao incorporar informações de lags, dias da semana e feriados institucionais e nacionais. Essa estrutura multivariada proporcionou um ajuste mais fino da série, capturando efeitos sazonais, eventos atípicos e comportamentos de consumo de curto prazo, tornando-se altamente aderente ao contexto decisório da empresa Segrob Notlad.

Em contrapartida, o modelo K-Nearest Neighbors (KNN) apresentou limitações significativas. Apesar de sua simplicidade e sensibilidade local, mostrou-se sensível a ruídos e à irregularidade da série, com desempenho inferior em relação às demais abordagens. Sua limitação em explorar dependências temporais profundas comprometeu sua precisão e aplicabilidade prática.

A avaliação dos modelos mais avançados, como Random Forest e Support Vector Regression (SVR), revelou diferenças importantes de comportamento. O Random Forest demonstrou maior capacidade de adaptação aos padrões sazonais e interações não lineares, sendo altamente eficaz em prever os picos de dezembro de 2024, especialmente nas semanas que antecedem o Natal. Já o SVR, embora sofisticado, não conseguiu captar adequadamente a variabilidade da série, apresentando métricas de erro elevadas e previsões excessivamente suavizadas.

Em termos de estratégia futura, recomenda-se: A exploração de modelos híbridos que combinam componentes sazonais e de tendência com algoritmos não lineares, a validação cruzada com múltiplas janelas temporais, aumentando a robustez das avaliações, a inclusão de variáveis externas adicionais (como clima, campanhas específicas, ou datas comemorativas setoriais) e a aplicação de técnicas de autoajuste de hiperparâmetros e monitoramento contínuo do desempenho preditivo, visando maior autonomia e responsividade do sistema.

Em suma, o estudo mostrou que modelos preditivos bem estruturados, com incorporação de conhecimento temporal e domínio operacional, oferecem ganhos concretos para o planejamento e a tomada de decisão no varejo de moda, permitindo antecipar demandas, otimizar estoques e alinhar ações promocionais com o comportamento real de consumo.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, M. M.; SILVA, T. A. da; SANTOS, J. C. dos. Metodologia CRISP-DM: potencialidades na descoberta do conhecimento em dados educacionais. *Revista Brasileira de Informática na Educação*, v. 28, n. 1, p. 109–131, 2020. Disponível em: <https://doi.org/10.5753/rbie.2020.28.01.109>. Acesso em: 20 jun. 2025.

COSTA, R. M.; LIMA, F. J. A.; VIEIRA, M. C. O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito. *Anais do*

Congresso Brasileiro de Sistemas, v. 10, p. 453–462, 2016. Disponível em: <https://www.example.com/artigo-crisp-credito.pdf>. Acesso em: 20 jun. 2025.

BORGES, Dalton. Aprendizado de Máquina – Análise Preditiva. Disponível em: daltonborges.com.br.

BORGES, Dalton. Regressão e Análise Preditiva – Árvores de Decisão. Disponível em: <https://daltonborges.com.br>.

BORGES, Dalton. Aprendizado de Máquina – Support Vector Machines. Disponível em: <https://daltonborges.com.br>.