UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE GESTÃO E NEGÓCIOS
GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO

LEANDRO RODRIGUES DE ÁVILA

Desenvolvimento de uma Dimensão de Produtos com Python: melhorando a qualidade da informação no setor varejista

ORIENTADOR: PROF. DR. JOSÉ EDUARDO FERREIRA LOPES

UBERLÂNDIA – MG
2024

LEANDRO RODRIGUES DE ÁVILA

Desenvolvimento de uma Dimensão de Produtos com Python: melhorando a qualidade da informação no setor varejista

Monografia apresentada ao Curso de Graduação em Gestão da Informação, da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de Bacharel.

Orientador: Prof. Dr. José Eduardo Ferreira Lopes

UBERLÂNDIA – MG

2024

RESUMO

Objetivou-se com este relato tecnológico descrever o processo de tratamento de uma base de dados de produtos para suportar decisões em uma empresa de comércio e distribuição. A situação problema identificada foi a dificuldade causada pela diversidade de nomenclaturas de produtos, que resultava em dados inconsistentes e falta de confiabilidade nas informações disponíveis. Como solução, foi adotado um processo de pré-processamento dos dados que incluiu a limpeza e normalização das informações, bem como a criação de uma base unificada utilizando códigos de barras como chaves únicas. Como resultados alcançados, destaca-se a melhoria significativa na qualidade dos dados, a agilidade no processo de categorização e a capacitação da empresa para tomar decisões informadas, resultando em uma melhor gestão do portfólio de produtos.

Palavras-chave: Varejo; Tratamento de Dados; Insights; Classificação.

SUMÁRIO

1. INTRODUÇÃO

O volume de dados gerado diariamente tem crescido de forma exponencial, e com isso, tem-se a necessidade de transformá-los em informações úteis para as empresas que buscam manter sua competitividade. Segundo Davenport e Harris (2007, p. 8), “a capacidade de uma organização de competir depende cada vez mais de sua habilidade de coletar, analisar e usar dados”. No ambiente de varejo, onde o fluxo de transações é constante, a capacidade de interpretar e tratar dados extraídos de diferentes fontes pode ser um diferencial estratégico (Davenport, Harris, 2007).

O setor de varejo, um dos mais dinâmicos e competitivos da economia global, caracteriza-se pela alta frequência e diversidade de transações, o que gera uma vasta quantidade de dados diariamente. De acordo com Levy e Weitz (2000, p. 23), “a complexidade do varejo moderno exige ferramentas analíticas robustas para lidar com um volume imenso de informações e para adaptar-se rapidamente às mudanças nas preferências dos consumidores”. Esses dados, quando corretamente tratados e analisados, podem revelar padrões importantes que ajudam as empresas a ajustar suas estratégias de marketing, otimizar a gestão de estoque e melhorar a eficiência operacional (Levy, Weitz, 2000).

Além disso, a análise dos dados de vendas também revela quais marcas e indústrias estão performando melhor no mercado. Isso permite às indústrias do setor identificar oportunidades de crescimento, avaliar o sucesso de seus produtos e ajustar suas estratégias de acordo com o comportamento do consumidor. A capacidade de analisar dados detalhados do desempenho de marcas e produtos específicos ajudam a entender as tendências de consumo e assim alinhar suas estratégias de acordo com a demanda do mercado.

Para lidar com esses desafios, o uso de notebooks em Python para o tratamento de dados tornou-se uma prática cada vez mais comum, devido à flexibilidade e ao poder de processamento que essa ferramenta oferece (McKinney, 2018). Ferramentas como Pandas e NumPy são frequentemente utilizadas para manipulação e análise de grandes volumes de dados, permitindo desde a padronização de formatos, tratamento de valores inconsistentes, até operações complexas de agregação e filtragem. Além disso, para a extração de dados de referência de fontes como PDF’s ou sites de código aberto, o uso

das ferramentas como Tabula e *WebScraping* do Python, facilita o processo de importação desses dados para compor tabelas “dimensões” da base de dados.

Além dessas bibliotecas, frameworks como Matplotlib e Seaborn são muito utilizados para a criação de visualizações iniciais que ajudam a identificar padrões e tendências nos dados. Essas visualizações auxiliam para o entendimento preliminar dos dados antes de sua integração em plataformas de Business Intelligence (BI) como o Power BI, onde são transformados em *dashboards* interativos. Estes *dashboards* por sua vez são utilizados para a comunicação visual de dados, oferecendo uma visão clara e objetiva que auxilia as organizações na tomada de decisões (Few, 2006).

No entanto, o processo de transformar dados brutos em informações úteis não é simples e requer uma abordagem estruturada que vai desde a coleta e preparação dos dados até a sua visualização. A limpeza dos dados, por exemplo, pode envolver a remoção de inconsistências, como valores ausentes ou duplicados, e a transformação de dados categóricos para formatos numéricos adequados, utilizando técnicas de engenharia de *features*. Croll e Yoskovitz (2014) destacam que “a interpretação eficaz dos dados requer um entendimento profundo tanto das técnicas analíticas quanto do contexto empresarial em que são aplicadas”, e é aí que a colaboração entre as áreas técnicas e estratégicas da empresa se torna vital.

Portanto, este relato tecnológico propõe-se a descrever o processo de tratamento de dados obtidos de notas fiscais de consumidores finais do varejo, utilizando notebooks em Python, melhorando consequentemente as visualizações criadas no Power BI. O objetivo é a otimização dos dados, que atualmente estão inconsistentes, oferecendo uma base de produtos (dimensão) para auxiliar a tomada de decisões no ambiente de varejo, fornecendo uma visão clara e baseada em dados concretos.

2 FUNDAMENTAÇÃO TEÓRICA

Os temas apresentados a seguir abordam a importância do tratamento e da análise de dados no contexto contemporâneo, com foco especial no setor de varejo. Em um cenário marcado pela explosão do volume de informações, que transformou o modo como empresas tomam decisões estratégicas.

2.1 A Era dos Dados e o Tratamento de Dados

Nos encontramos em uma era marcada pela massiva geração de dados, fenômeno que transformou o panorama econômico e social, inaugurando o que Diebold (2012) chama de "Era dos Dados". Essa nova era é caracterizada pela produção constante de grandes volumes de dados provenientes de diversas fontes, como transações comerciais, interações nas redes sociais, dispositivos conectados e outros sistemas digitais. Esse aumento exponencial no volume de dados impulsionou a adoção de técnicas de coleta, tratamento e análise, capazes de converter esses dados em informações para a tomada de decisões no mundo dos negócios (Diebold, 2012).

A evolução da tecnologia, com o advento de sistemas de computação em nuvem e o desenvolvimento de algoritmos de aprendizado de máquina, permitiu que os dados passassem de um mero conjunto de registros a um ativo importante para as organizações. No entanto, como apontam Kimball e Ross (2013), o simples fato de acumular dados não é suficiente. É necessário desenvolver um processo estruturado de tratamento e análise que garanta a qualidade e a utilidade das informações extraídas. Neste contexto, o tratamento de dados envolve diversas atividades como a limpeza, padronização, integração e transformação de dados, que têm como objetivo assegurar que os dados brutos se tornem confiáveis para a geração de *insights* (Kimball; Ross, 2013).

O tratamento de dados, conforme destacado por Davenport e Harris (2007) é um desafio que envolve técnica e estratégia, eles afirmam que a capacidade de uma organização de competir está diretamente relacionada à sua habilidade de coletar, analisar e usar dados. As empresas que implementam processos robustos de tratamento de dados são capazes de identificar padrões comportamentais dos consumidores, prever tendências de mercado e ajustar suas operações com mais rapidez e precisão. A falta de tratamento adequado, por outro lado, pode resultar em decisões mal-informadas, que impactam negativamente a eficiência operacional e a satisfação dos clientes (Davenport; Harris, 2007).

McKinney (2018) observa que o Python se tornou uma das ferramentas mais eficazes para o tratamento de grandes volumes de dados. Com suas diversas bibliotecas que auxiliam em praticamente todas as etapas de ETL. O uso dessas ferramentas se consolidou em diversas indústrias, incluindo o varejo, dada a sua capacidade de lidar com bases de dados extensas e diversificadas (McKinney, 2018).

Contudo, é notável que a "Era dos Dados" trouxe consigo novos desafios. A transformação de dados brutos em *insights* acionáveis requer ferramentas funcionais e

simples de utilizar, além de uma compreensão profunda do contexto no qual esses dados estão inseridos. O sucesso das empresas na era digital depende, em grande medida, de sua capacidade de navegar por este cenário complexo e em constante evolução (Fallet; Isolani, 2022).

2.2 O Papel dos Dados no Setor de Varejo

O setor de varejo, por sua própria natureza, é um dos principais geradores de dados no ambiente empresarial contemporâneo. Cada transação realizada produz uma série de informações que podem ser exploradas para obter *insights* sobre o comportamento do consumidor, o desempenho de produtos, a eficiência de operações e a eficácia das campanhas de marketing. Como destacam Levy e Weitz (2000), o varejo moderno é caracterizado pela elevada diversidade e frequência de transações, o que o torna um ambiente particularmente propício para a coleta e análise de dados.

O uso eficaz de dados no setor de varejo pode dizer muito sobre as preferências e padrões de comportamento dos consumidores, dados transacionais podem revelar, por exemplo, variações sazonais na demanda por determinados produtos, tendências de consumo em diferentes regiões e padrões de recompra. Essas informações permitem que as empresas ajustem suas estratégias de marketing e desenvolvimento de produtos com maior precisão (Gonçalves; Lima, 2021). Conforme apontam Kotler e Keller (2012), a capacidade de entender e antecipar as necessidades dos consumidores é uma das chaves para o sucesso no varejo, e a análise de dados desempenha um papel central nesse processo.

Além disso, a análise detalhada dos dados de vendas para o consumidor final contribui para o sucesso das indústrias que operam em ambientes de alta competitividade, como o varejo. Ao monitorar continuamente o desempenho de seus produtos, as indústrias podem identificar oportunidades de crescimento, bem como áreas que necessitam de ajustes (Levy; Weitz, 2000).

No entanto, para que esses resultados sejam alcançados, é necessário que as empresas do setor de varejo adotem boas práticas de tratamento e análise de dados. Como argumentam Davenport e Harris (2007), o sucesso da análise de dados depende da qualidade das informações coletadas e da capacidade das empresas de transformar esses dados em informações acionáveis. No caso do varejo, isso significa integrar dados

de múltiplas fontes como sistemas de pontos de venda, sites ou catálogos de produtos, e garantir que esses dados sejam limpos, consistentes e atualizados.

Assim, fica nítido que o papel dos dados no setor de varejo está ligado a otimização de operações. A análise de dados permite que as empresas identifiquem tendências de consumo, personalizem suas ofertas e maximizem a eficiência de suas operações. Com a digitalização do varejo e a integração de múltiplos canais de venda é necessário o uso estratégico dos dados disponíveis. A visualização adequada desses dados, por meio de *dashboards* e relatórios interativos, pode ajudar os gestores a tomar decisões mais informadas e ágeis, garantindo que suas estratégias estejam sempre alinhadas às necessidades do mercado (Few, 2006).

2.3 Ferramentas para Tratamento dos Dados

No cenário contemporâneo, o volume crescente de dados exige ferramentas para garantir que essas informações possam ser devidamente processadas e analisadas. O tratamento de dados é um processo que envolve a coleta, limpeza, transformação e análise dos dados. A utilização de ferramentas adequadas para cada etapa desse processo pode determinar a precisão das análises realizadas (Kimball; Ross, 2002).

A linguagem de programação Python tem se destacado como ferramenta para o tratamento de dados devido à sua flexibilidade e à riqueza de suas bibliotecas. A biblioteca Pandas, desenvolvida por McKinney (2018), é amplamente reconhecida por sua capacidade de manipular e analisar dados tabulares. A Pandas fornece estruturas de dados, como DataFrames, que facilitam a realização de operações complexas, como filtragem, agregação e transformação dos dados. McKinney (2018) observa que "Pandas oferece uma abordagem intuitiva para carregar, manipular e analisar dados, tornando-se uma ferramenta muito usada por cientistas de dados e analistas".

Complementando o Pandas, com a biblioteca NumPy é possível realizar operações numéricas e o gerenciamento de *arrays* multidimensionais. "NumPy proporciona suporte para cálculos matemáticos avançados e é frequentemente utilizado em conjunto com outras bibliotecas, como Pandas, para realizar operações matemáticas e estatísticas complexas." Essa combinação permite a manipulação eficiente de grandes volumes de dados e a execução de análises matemáticas sofisticadas (Oliphant, 2006).

Outras bibliotecas que são importantes mencionar, pois foram bastante utilizadas neste projeto é a Tabula e a *BeautifulSoup*, O Tabula é útil principalmente para importar

dados que estão em formatos de arquivo não estruturados, como PDFs de notas fiscais e relatórios. A sua integração com Python permite que os dados sejam extraídos e convertidos em formatos utilizáveis para tratamentos e análises subsequentes (Rechner, 2016). O *BeautifulSoup* também utilizado para extração de dados, com viés para processos de *Webscraping*. O algoritimo facilita a extração de informações de arquivos HTML e XML, possibilitando a coleta de dados de websites de forma eficiente e estruturada. A sua simplicidade combina bem com Python, permitindo realizar operações de *scraping* de maneira ágil e flexível (Mitchell, 2015).

Após o tratamento inicial dos dados, é preciso garantir que os dados estejam limpos e consistentes. Isso pode envolver a remoção de duplicatas, o preenchimento de valores ausentes, padronização de campos e a correção de erros de entrada. Ferramentas e técnicas de limpeza de dados asseguram a qualidade e a precisão das análises. Como Kimball e Ross (2013) ressaltam, "a qualidade dos dados é um componente essencial para a eficácia da análise de dados, e a limpeza apropriada é necessária para garantir que as conclusões sejam precisas e baseadas em informações corretas".

2.4 Visualização de Dados e Sua Importância

A visualização de dados é uma etapa crítica no processo de análise, permitindo que os analistas interpretem e compreendam grandes volumes de informações de maneira intuitiva. A capacidade de criar representações visuais dos dados ajuda a identificar padrões, tendências e *insights* que podem não ser imediatamente evidentes através da análise numérica isolada (Kirk, 2016).

A biblioteca Matplotlib é amplamente utilizada para a criação de gráficos e visualizações em Python. Segundo Hunter (2007), "Matplotlib fornece uma plataforma flexível e poderosa para a criação de gráficos estáticos, animados e interativos." Ela permite o controle detalhado sobre a aparência dos gráficos, possibilitando a personalização completa para atender às necessidades específicas de visualização. A Matplotlib é frequentemente usada em combinação com outras bibliotecas para criar gráficos complexos e informativos (Hunter, 2007).

Seaborn, por outro lado, é uma biblioteca que praticamente complementa o Matplotlib e oferece uma interface de alto nível para a criação de visualizações estatísticas. De acordo com Waskom et al. (2018), "Seaborn é projetado para facilitar a criação de gráficos estatísticos complexos e esteticamente agradáveis." A combinação

de Seaborn com Pandas permite a criação de visualizações detalhadas e informativas que ajudam a comunicar os resultados das análises de dados de forma eficaz (Waskom, 2018).

Além das bibliotecas de visualização em Python citadas, as ferramentas de Business Intelligence (BI) como o Power BI são utilizadas para integrar e apresentar os dados de forma interativa. O Power BI oferece uma plataforma para criar dashboards que consolidam informações de diferentes fontes e apresentam uma visão consolidada e dinâmica dos dados. Conforme Few (2006), "as ferramentas de BI transformam dados em visualizações interativas que facilitam a interpretação e a tomada de decisões baseadas em dados concretos."

Evidentemente, a visualização eficaz melhora a compreensão dos dados e facilita a comunicação para diferentes áreas interessadas dentro da organização. Sua utilidade é reconhecida tanto na análise exploratória como na apresentação dos resultados finais analisados. Apresentar os dados de forma clara e acessível ajuda a garantir que os *insights* possam ser compreendidos e utilizados para tomar decisões informadas. A combinação de ferramentas de visualização e BI proporciona uma comunicação efetiva dos dados e para a tomada de decisões baseada em informações precisas e relevantes (Few, 2012).

2.5 Machine Learning e Inteligência Artificial

A integração de técnicas de Machine Learning (ML) e Inteligência Artificial (IA) tem transformado a forma como os dados são analisados e utilizados. Machine Learning refere-se a um conjunto de técnicas e algoritmos que permitem que sistemas aprendam a partir de dados e façam previsões ou tomem decisões com base em padrões identificados. A IA, por sua vez, abrange uma gama mais ampla de tecnologias que visam simular aspectos da inteligência humana, como o reconhecimento de padrões, a tomada de decisões e a interação em linguagem natural (Jordan; Mitchell, 2015).

No contexto do varejo, o uso de ML e IA tem se mostrado técnicas importantes para a análise e interpretação de grandes volumes de dados. Modelos de aprendizado supervisionado, como regressão e classificação, são frequentemente empregados para prever comportamentos futuros dos consumidores, como a propensão à compra e a retenção de clientes, além de contribuir para classificar dados internos. Esses modelos

utilizam dados históricos para treinar algoritmos que, posteriormente, fazem previsões sobre novos dados (Kumar; Rajeevan, 2020).

De acordo com Provost e Fawcett (2013), “Machine Learning fornece ferramentas e técnicas para aprender a partir de dados, detectar padrões e fazer previsões que são fundamentais para a tomada de decisões baseada em dados. ”. Uma aplicação, por exemplo, de ML no varejo é a classificação de dados de compra, pois por ter milhões de dados entrando a cada hora, muitas descrições de produtos seguem diferentes padrões, então o algoritmo consegue identificar esses padrões para classificar rapidamente todos produtos em apenas uma descrição pré-estabelecida. Além disso, a IA é utilizada para automação e otimização de processos como gestão de estoque e previsão de demanda, por exemplo (Peterson; McFarlane, 2021).

Todos os fundamentos apresentados acima evidenciam a importância de uma abordagem sistemática e bem estruturada no tratamento e análise de dados no setor de varejo. O avanço das tecnologias e ferramentas de análise de dados, incluindo Python e outras ferramentas de visualização de dados, tem possibilitado um entendimento mais preciso das operações comerciais e das preferências dos consumidores se tornando uma estratégia trivial em mercados com alta competitividade.

A análise detalhada dos dados e a utilização de ferramentas apropriadas melhoram a eficiência operacional e consequentemente ajuda a orientar as estratégias de negócios e aprimorar a experiência do cliente. Portanto, a adoção dessas tecnologias e métodos é cada vez mais comum pelas empresas que buscam se destacar no dinâmico e competitivo ambiente do varejo, maximizando o valor dos dados e ajustando suas estratégias conforme as tendências e comportamentos deles (Bellini; Lam; Turrini, 2020).

3 CONTEXTO INVESTIGADO E SITUAÇÃO PROBLEMA

O presente estudo foi realizado em uma das maiores empresas de comércio e distribuição de Minas Gerais, que se destaca pela sua vasta gama de produtos e pela complexidade de suas operações. A empresa, ao lidar diariamente com um volume significativo de dados provenientes de notas fiscais de pequeno e médio varejista, implementou um algoritmo de Machine Learning (ML) para classificar produtos com base em suas descrições. Este algoritmo utiliza uma base de treinamento que permite identificar similaridades entre as descrições dos produtos e classificá-los em categorias

predefinidas, além disso a partir da descrição também são extraídas, através de técnicas de *regex* informações como unidade de medida (GR/ML/UND), assim como valor da medida identificada.

Entretanto, a diversidade de nomenclaturas utilizadas por diferentes varejistas para se referir ao mesmo produto, gera uma série de desafios, inclusive para o uso de expressões regulares. Cada estabelecimento pode nomear o mesmo item de maneira distinta, de modo que o mesmo item pode ter muitas descrições diferentes (Tabela 1) e também descrições parecidas ou até iguais, podem se tratar de itens distintos, resultando em classificações ambíguas e consequentemente, erradas. Essa inconsistência compromete a precisão das informações e gera duplicidade de descrições para o mesmo produto, aumentado desnecessariamente a dimensão de produtos e dificultando a análise e a tomada de decisões estratégicas.

Tabela 1 – Produtos com descrições diferentes tendo suas classificações afetadas

CODIGO DE BARRAS	PRODUTO	MARCA	FABRICANTE	CATEGORIA
7896213002503	ROSQUINHAS SABOR COCO VITARELLA	None	None	BISCOITO
7896213002503	ROSQUINHA VITARELLA COCO 350G	Vitarella	M DIAS BRANCO	BISCOITO
7896213002503	ROSQUI VITARELLA DE COCO 350G	VITARELLA	M DIAS BRANCO	ACUCAR
7896213002503	ROSQUINHA GRANEL COCO VITAREL	None	None	BISCOITO
7896213002503	ROSQ VITARE COCO 350	None	None	BISCOITO

Fonte: Autor (2024)

A literatura destaca a importância de uma abordagem sistemática na aplicação de modelos de Machine Learning. Segundo Alpaydin (2010), “o aprendizado de máquina é uma subárea da inteligência artificial que se concentra em desenvolver algoritmos que permitem que os computadores aprendam a partir de dados”. No entanto, a eficácia desses modelos é afetado pela qualidade dos dados utilizados para o treinamento. A presença de dados ambíguos e mal classificados pode levar a resultados imprecisos, comprometendo a capacidade da empresa de responder rapidamente às mudanças nas condições do mercado (Alpaydin, 2010).

Além disso, a implementação de algoritmos de ML requer um entendimento profundo das características dos dados e das variáveis que influenciam as classificações. De acordo com Jain et al. (2016), “a integração de técnicas de aprendizado de máquina com uma análise cuidadosa dos dados pode resultar em melhorias significativas na eficiência operacional”. Contudo, para que essa transformação ocorra de maneira eficaz,

é necessário que a empresa desenvolva um processo mais assertivo de tratamento e normalização dos dados, garantindo que as informações sejam consistentes e confiáveis (Jain, et al., 2016).

Diante desse cenário, a situação problema se configura na necessidade de otimizar o processo de classificação de produtos, minimizando as inconsistências e melhorando a precisão das informações. A adoção de técnicas de pré-processamento de dados, como a normalização de descrições ou identificação de campos “chave”, antes de partir para a utilização de algoritmos de aprendizado, pode ser uma solução viável para enfrentar os desafios apresentados.

4 INTERVENÇÃO ADOTADA

A proposta para solucionar o problema consiste na criação de uma dimensão de produtos, com o objetivo de otimizar a classificação e categorização dos produtos. A estratégia adotada foi descoberta através da exploração dos dados e identificação de um campo “chave”, ou seja, que é único para cada produto, no caso o código de barras como pode ser visto na Tabela 1. Porém, é preciso ressalvar que alguns produtos são enviados sem código de barra, nesses casos a melhor solução seria melhorar a base de treinamento do algoritmo de Machine Learning já utilizado ou encontrar um algoritmo mais eficiente. Hoje aproximadamente 65% dos produtos são enviados com o código podendo então ser classificados a partir deles, gerando uma melhora significativa para a base geral.

A padronização prévia através do código de barras, caso ele seja enviado, visa evitar duplicidades e inconsistências, garantindo uma base de referência para os novos produtos que integram diariamente as tabelas fato. Para construir essa dimensão de produto, foi necessário integrar três principais fontes de dados. Primeiramente, utilizou-se própria base de dados da empresa no Databricks, que possibilita a extração dos códigos de barras mais recorrentes e assim definir suas respectivas características, que uma vez definidos os valores serão padronizados para todos os produtos que entrarem na base com o respectivo código.

Em paralelo, foram extraídos dados diretamente dos catálogos de produtos de fabricante e distribuidores. Através do uso do Tabula, ferramenta eficaz na extração de dados de documentos PDF, foi possível coletar detalhes sobre os produtos, como descrições, variações de embalagem, peso, e outras características que complementam a

base interna desenvolvida. Essa integração ajudou garantir a completude das informações na dimensão de produto, assegurando a qualidade e confiabilidade dos dados utilizados na classificação.

Além dessas duas fontes, foi realizada a extração automatizada de informações de sites de distribuidoras por meio de técnicas de *WebScraping*. Essa abordagem permitiu capturar dados atualizados diretamente de sites de distribuidores tanto da própria empresa como dos concorrentes. A coleta recorrente dessas informações possibilita a atualização dinâmica da dimensão de produto, garantindo que as análises estejam sempre alinhadas às mudanças do mercado.

Vale destacar que os dados extraídos de fontes externas como catálogos e sites, originalmente não se encontravam no formato correto da dimensão criada, para ajustar esse formato foram usadas bibliotecas do Python como Pandas e Numpy. Dessa forma os dados extraídos de diferentes fontes apresentavam em sua versão final o mesmo formato, facilitando a integração dos mesmos.

A integração dessas três fontes permitiu a criação de uma base única de produtos, que facilitou a classificação dos produtos de maneira rápida e precisa, evitando erros e inconsistências que anteriormente comprometiam a confiabilidade dos dados. Com a utilização do campo código de barras como referência principal, a intervenção proposta teve como objetivo de tornar o processo de categorização mais eficiente e assertivo, contribuindo diretamente para a qualidade das análises e decisões posteriores da empresa.

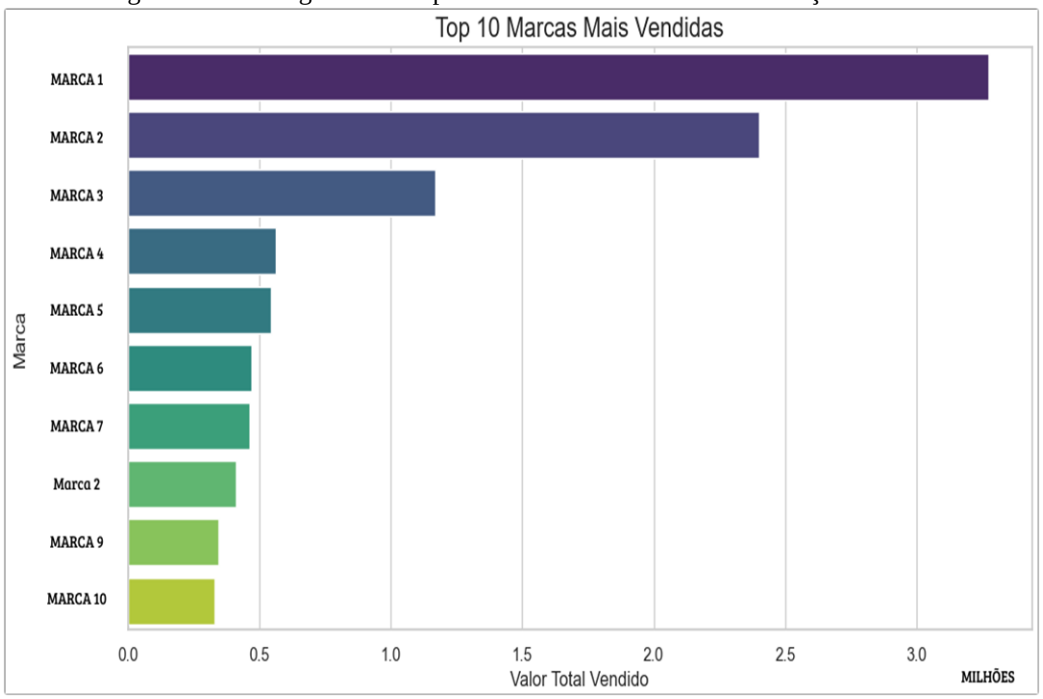
5 RESULTADOS ALCANÇADOS

A intervenção proposta está sendo implementada atualmente como uma das demandas prioritárias a empresa, o processo de mapear produtos do varejo é desafiador, basta olhar para gama de produtos dentro de um supermercado. Porém mesmo estando no começo da implementação, resultados significativos já foram obtidos, principalmente no que se refere à eficiência e precisão na classificação dos produtos. A construção de uma dimensão de produto, centrada no código de barras, trouxe uma grande melhora na qualidade dos dados, resultando em uma base confiável para as operações. A padronização dos produtos utilizando esse código único eliminou duplicidades e inconsistências, garantindo uniformidade nas informações.

Um dos principais resultados foi o aumento na agilidade de categorização dos produtos. A integração das três fontes de dados — a base interna no Databricks, catálogos de fabricantes extraídos com o Tabula, e informações de sites de distribuidoras capturadas via WebScraping — automatizou a classificação, reduzindo significativamente o tempo necessário para processar e analisar os dados. Antes, a categorização dependia de uma alta intervenção manual para corrigir erros do classificador, o que gerava atrasos e maiores chances de erros. Agora, o processo é mais ágil e eficiente.

Outro resultado que vale a pena destacar foi a corretude aumentada nos relatórios e análises gerados, principalmente quando as informações de produtos são ligadas aos KPI's (Key Performance Indicator), como preço unitário e quantidade vendida. A eliminação de dados incorretos e duplicados permitiu que os relatórios criados no Power BI refletissem de maneira correta a situação operacional dos clientes, facilitando a tomada de decisões, tornando então a solução mais atrativa e viável para eles.

Figura 1 – Ranking de marcas por valor vendido antes da intervenção adotada

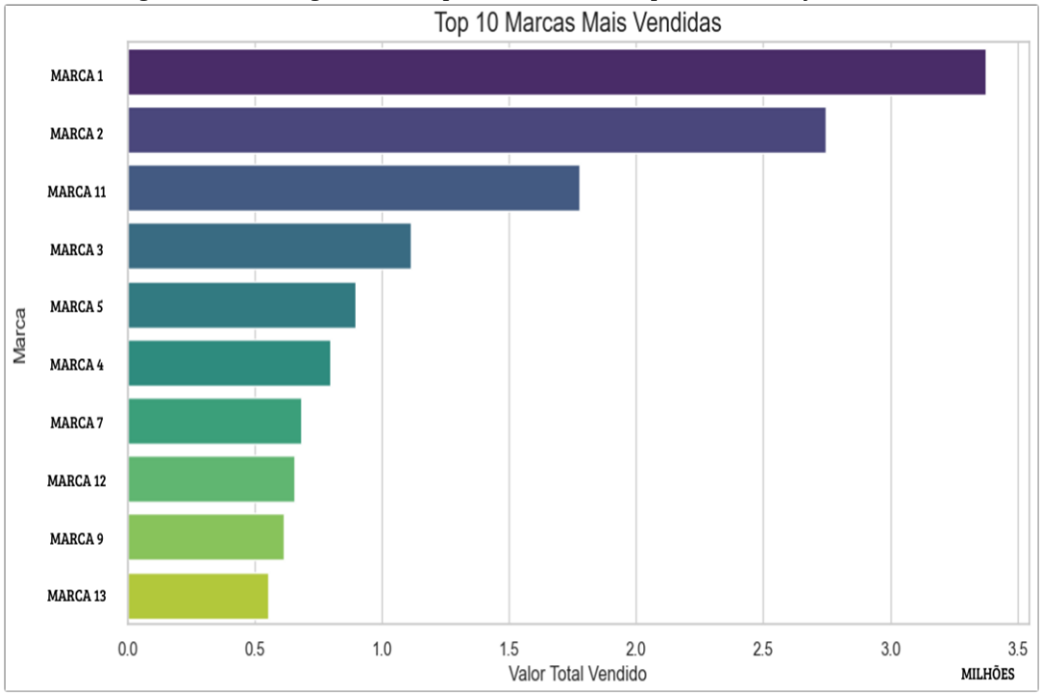


Fonte: Autor (2024)

Como foi evidenciado na Figura 1 por conta da falta de padronização das descrições, algumas marcas acabam duplicando por estarem escritas de forma diferente,

por exemplo as posições 2 e 8 do ranking que aparecem como “MARCA 2” e “Marca 2”, mesmo se tratando da mesma marca.

Figura 2 – Ranking de marcas por valor vendido após a intervenção adotada



Fonte: Autor (2024)

Comparando a Figura 2 com a Figura 1, foi visto algumas diferenças significativas no ranking de marcas. Primeiro os valores aumentaram para todas presentes entre as dez mais vendidas, que se deu provavelmente por conta de valores antes sem classificação (nulos) que passaram a ser identificadas após a implementação. Além disso algumas marcas deixaram de fazer parte da lista (“MARCA 6”, “Marca 2”, “MARCA 10”), enquanto outras passaram a compor a mesma (“MARCA 11”, “MARCA 12”, “MARCA 13”). Por último houve troca de posições no ranking, o que é um fator muito relevante para análises, por exemplo, a “MARCA 3” que antes ocupava a terceira posição foi para quarta, enquanto a “MARCA 11” que antes nem aparecia no ranking passou a ocupar a terceira posição. Essa comparação evidenciou um dos resultados obtidos pela solução, que trouxe uma melhora para as análises mais corretas.

Até o momento, a base de dados construída contém mais de 20 mil códigos de barras únicos, com suas respectivas classificações mapeadas, abrangendo mais de 18 categorias de produtos. Esse volume expressivo de informações estruturadas está permitindo uma visão ampla e detalhada do portfólio de produtos, com impacto direto no produto oferecido pelo projeto.

6 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo descrever o processo de tratamento de dados oriundos de notas fiscais de consumidores do varejo, utilizando ferramentas como notebooks em Python para transformar dados brutos em informações relevantes levadas através de visualizações com Power BI, ajudando na tomada de decisões. O objetivo principal foi o desenvolvimento de uma base de produtos confiável (dimensão) que auxilie na padronização e classificação de produtos, proporcionando maior precisão e agilidade no ambiente de varejo.

Ao longo do projeto, alcançou-se um progresso relevante, se comparado com o problema inicial encontrado, com a criação de uma base contendo mais de 20 mil códigos de barras únicos, distribuídos em mais de 18 categorias de produtos. Pode-se considerar que esse resultado atendeu às expectativas iniciais e destacou a importância de trabalhar com dados estruturados e bem classificados para melhorar a competitividade empresarial, como discutido por autores como Davenport e Harris (2007) e Levy e Weitz (2000).

Apesar de a dimensão de produtos ainda estar em processo de construção, já foi possível disponibilizar soluções viáveis para indústrias e clientes específicos. Um exemplo é uma grande empresa do setor de artigos de papelaria e produtos de uso diário, para a qual foi entregue um produto bem classificado e adaptado às necessidades específicas, como comparação de preço médio e faturamento de suas categorias com a de seus concorrentes. Esse resultado mostra que, mesmo em fases intermediárias do projeto, a base de dados já se apresenta como um recurso importante para a empresa.

O processo de desenvolvimento da base enfrentou desafios, como a integração de dados provenientes de diferentes fontes (bases internas, catálogos de fabricantes e dados obtidos via WebScraping), além da manipulação de grandes volumes de informações. Essas dificuldades, no entanto, trouxeram oportunidades de novos aprendizados em ferramentas analíticas, como Pandas, NumPy e Matplotlib, que foram importantes para chegar aos resultados alcançados.

Assim, este relato tecnológico cumpriu seu papel ao descrever detalhadamente as etapas do projeto, desde o tratamento até a consequente visualização dos dados em *dashboards* interativos no Power BI. A otimização dos dados, antes inconsistentes,

agora oferece uma base de produtos pré-definida, capaz de fornecer uma visão clara e concreta para o apoio à tomada de decisões no setor de varejo. Esse projeto também reforça a importância da colaboração entre áreas técnicas e estratégicas dentro das organizações, alinhando-se às necessidades empresariais e ao uso eficaz de técnicas analíticas (Croll e Yoskovitz, 2014).

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALPAYDIN, E. **Introduction to Machine Learning.** 3. ed. Cambridge: MIT Press, 2010.

BELLINI, H.; LAM, T.; TURRINI, V. Data-Driven Retail: Unlocking Business Insights with Advanced Analytics. **Journal of Retail Innovation**, v. 12, p. 88-102, 2020.

CROLL, A.; YOSKOVITZ, B. **Lean Analytics: Use Data to Build a Better Startup Faster**. O'Reilly Media, 2014.

DAVENPORT, T. H.; HARRIS, J. G. **Competing on Analytics: A Nova Ciência da Competição**. Harvard Business School Press, 2007.

DIEBOLD, F. X. The Era of Big Data. **The Journal of Economic Perspectives**, 26(2), p. 1-18, 2012.

FEW, S. **Information Dashboard Design: A Comunicação Visual Eficaz de Dados**. O'Reilly Media, 2006.

FEW, S. **Show Me the Numbers: Designing Tables and Graphs to Enlighten**. Analytics Press, 2012.

GONÇALVES, R.; LIMA, A. Análise de Dados no Varejo: Estratégias e Benefícios. **Revista Brasileira de Gestão e Negócios**, 2021.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90-95, 2007.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. **ACM Computing Surveys**, v. 31, n. 3, 1999.

JORDAN, M. I.; MITCHELL, T. M. Machine Learning: Trends, Perspectives, and Prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. Wiley, 2013.

KIMBALL, R. **The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence**. Wiley, 2013.

KOTLER, P.; KELLER, K. L. **Administração de Marketing**. 14. ed. São Paulo: Pearson Prentice Hall, 2012.

KUMAR, V.; RAJEEVAN, N. P. Machine Learning and Artificial Intelligence Applications in Retail. **International Journal of Advanced Science and Technology**, v. 29, n. 3, p. 123-137, 2020.

LEVY, M.; WEITZ, B. A. **Administração de Varejo**. São Paulo: Atlas, 2000.

MCKINNEY, W. **Python para Análise de Dados: Manipulação de Dados com Pandas, NumPy e IPython.** 2. ed. São Paulo: Novatec Editora, 2018.

MITCHELL, R. **Web Scraping with Python: Collecting More Data from the Modern Web.** O'Reilly Media, 2015.

OLIPHANT, T. E. **A Guide to NumPy**. Trelgol Publishing, 2006.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**. O'Reilly Media, 2013.

RICCI, F.; et al. **Recommender Systems Handbook**. Springer, 2011.

RECHNER, T. Practical Web Scraping for Data Science. **Packt Publishing**, 2016.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. Pearson, 2010.

WASKOM, M.; et al. Seaborn: Statistical Data Visualization. **Journal of Open Source Software**, v. 6, n. 60, 3021, 2018.