Supermarket DataInsights:

From Raw Data to Actionable Insights

Presented by
Caio Alcântara

7/5/2025

Outline

Executive Summary

This report presents a comprehensive data analytics and forecasting project designed to analyze and predict supermarket product sales performance. The primary objective was to transform raw transactional data into actionable insights that support informed business decision-making and strategic planning.

The project utilized a dataset containing historical sales records exported from the company's transaction systems, covering the period from April 2024 through April 2025. These records include detailed information such as product codes and descriptions, quantities sold, unit costs, total revenue, and profit margins.

The workflow encompassed several key phases. First, data wrangling processes were implemented to clean numeric fields, standardize formats, and translate product descriptions for consistency. Next, exploratory data analysis was conducted to identify patterns in sales volume, revenue trends, and monthly profitability across different product categories.

A supervised machine learning model was then developed using linear regression to forecast total monthly profit based on sales volume, average unit cost, unit profit, and margin percentage. Model performance was evaluated using mean absolute error and R^2 score, providing a quantitative assessment of prediction accuracy.

Executive Summary

As part of the exploratory analysis, the project also introduced a segmentation framework that classified products into four performance tiers High Performer, Medium-High, Medium-Low, and Low Performer, based on the average of key metrics such as units sold, total profit, total sales, and margin percentage. This categorization helped highlight differences in product behavior and supported more intuitive visual comparisons across items.

To ensure results were accessible to stakeholders, an interactive dashboard was created with Streamlit. This dashboard enables users to filter data by product and date range, view key performance indicators, explore time series trends, and compare predicted versus actual profitability metrics.

Overall, the project demonstrates a complete end-to-end data science pipeline, combining robust data preparation, predictive modeling, and intuitive visual reporting. The findings offer clear insights into sales dynamics and profitability drivers, supporting data-driven decisions to optimize product strategy and maximize revenue.

Introduction

Supermarkets operate in a highly competitive environment where understanding product sales performance and profitability trends is essential to sustaining growth and optimizing operations. Every day, large volumes of transactional data are generated, capturing detailed records of product-level sales, costs, and margins. While this information can offer valuable insights into revenue drivers and customer behavior, it often remains underutilized without systematic analysis and clear reporting.

This project was developed to address this challenge by applying modern data analytics and forecasting techniques to supermarket sales records covering a twelve-month period from April 2024 through April 2025. By transforming raw transaction data into structured insights, the project aims to support evidence-based decision-making and help managers identify the factors that most influence financial outcomes.

Beyond describing historical performance, the analysis explores how sales patterns and profitability evolve over time, highlighting differences among products and categories. It also introduces a segmentation approach to classify products into performance tiers, helping to benchmark relative contribution and inform product strategy. Finally, a predictive modeling component provides estimates of future monthly profit, adding a forward-looking perspective that can support planning and resource allocation.

Introduction

Specifically, this report seeks to answer the following business and operational questions:

- How do factors such as units sold, unit cost, and margin percentage vary across different products?
- Can we develop a predictive model capable of estimating future monthly profit based on historical performance data?
- How can products be categorized into performance tiers to better understand their relative contribution and quality?
- Which product categories and individual items contribute most significantly to total sales revenue and profitability?
- What visual tools can help stakeholders explore trends and make evidence-based decisions?

By addressing these questions, the project demonstrates the value of data-driven approaches in retail. It shows how structured analysis, predictive modeling, and interactive visualization can improve transparency, guide strategy, and strengthen financial performance in a competitive market.

Methodology

Methodology Outline

Data Collection and Preparation

The dataset used in this project was sourced from Linear Systems, the supermarket's internal transaction management platform. Monthly sales records were exported in CSV format, covering the period from April 2024 to April 2025. The raw dataset included product-level information such as product codes and descriptions, transaction dates, quantities sold, unit prices, and profit-related metrics.

The data preparation process included the following key steps:

1. **Data Importation:** The raw CSV file was loaded into a pandas DataFrame for processing. Initial inspection identified formatting inconsistencies, particularly in numeric fields.

2. **Numeric Field Cleaning:** Several columns used commas as decimal separators due to locale-specific formatting (e.g., "3,50" instead of "3.50"). These were replaced with periods and converted to proper float types for calculation purposes.

3. **Missing Value Handling:** Rows containing missing or non-numeric values in critical fields (such as unit cost, total profit, or margin percentage) were removed to ensure data integrity.

Data Collection and Preparation

4. **Type Conversion:** All numeric fields were explicitly converted to appropriate numerical types (e.g., float64) to support reliable aggregation and modeling later in the workflow.

5. **Product Description Translation and Validation:** Product descriptions, originally recorded in Portuguese, were translated into English to maintain clarity and consistency in reporting. Additionally, a frequency chart of the most common description words was generated to help identify possible spelling inconsistencies, duplicate variants, or formatting issues in product names. This visual check supported targeted corrections and improved the uniformity of labels used throughout the analysis.

6. **Field Renaming and Formatting:** Column names were renamed to standardized, English-based labels with clear formatting. For instance, “Lucro Total” became “Total_Profit”, and “Margem %” became “Margin%”. This improved readability and ensured compatibility with visualization tools.

7. **Validation:** After cleaning, summary statistics were generated to validate data ranges and confirm the presence of expected values in key columns. A final visual inspection helped identify any lingering anomalies.

Data Collection and Preparation

- The data preparation process combined meticulous cleaning, translation, and validation to ensure a reliable, consistent, and analysis-ready dataset.

- Locale-specific numeric formats were standardized by converting European-style decimals to the common period format. Product descriptions were translated from Portuguese to English and audited through frequency charts to detect and correct inconsistencies such as typos or duplicate variants, improving label uniformity.

- Critical numeric fields were explicitly type-cast, and rows with missing or invalid values in key columns were removed to maintain data integrity.

- These combined steps ensured a high-quality, reproducible dataset suitable for robust exploratory analysis and modeling, reflecting strong attention to detail and real-world business applicability.

Exploratory Data Analysis (EDA)

The exploratory analysis presented in this project focuses on a single-month snapshot (March 2025), offering a product-level view of sales and profit dynamics. This example demonstrates the methodology applied to one month but can be adapted to any other period by adjusting the filtering criteria.

Rather than exploring long-term trends or seasonality, the goal was to extract insights for the selected month, identifying key revenue drivers, profit patterns, and product segmentation in that specific context.
The analysis followed these steps:

1. **Sales and Profit Distribution by Product:**
   Products were ranked by their sales and profit, identifying both top sellers and low-performing items. This provided a clear view of which products had the most commercial impact in the selected month.

2. **Margin % Variability**
   A boxplot of Margin % allowed the detection of pricing consistency and anomalies, including products with negative margins. This helped assess the supermarket's profitability profile for April.

Exploratory Data Analysis (EDA)

3. **Product Segmentation Snapshot**
   Based on their contribution to sales, the products were segmented into categories A, B, C and D. This revealed a familiar Pareto distribution, in which a small number of products generated the majority of sales.

Key Takeaways

This analysis represents an example month and can be replicated in other periods to monitor shifts in product performance, pricing strategies, and sales concentration over time.

Actionable Insights

The EDA findings suggest focusing marketing and inventory resources on top sellers, optimizing mid-tier products through promotions or bundling, and addressing underperformers by cost reduction, price adjustments, or phasing them out.

These data-driven actions can improve overall profitability and operational efficiency, ensuring resources are directed where they have the most impact.

Segmentation

The product segmentation process aimed to classify the supermarket's items into distinct performance groups, enabling a more strategic focus on product management. Rather than applying complex clustering models, a straightforward approach was used, combining business logic with data-driven metrics.

Instead of relying solely on cumulative sales, the segmentation categories were defined using a normalized average of the main performance indicators, including Total Sales, Total Profit, and Amount Sold. Each product's performance across these metrics was scaled to ensure comparability and then averaged to form a balanced view of its overall contribution.

Products were then ranked according to this combined score, and assigned to four categories:

- **Category A:** The top-performing products, representing approximately 2.6% of the catalog (~59 products). Despite their small number, these items generated the majority of the supermarket's revenue and profit, making them critical to business success.

- **Category B and C:** These middle-tier products contributed moderate sales volumes, offering opportunities for optimization through pricing, promotion, or stock management.

Segmentation

- **Category D:** The largest category in terms of product count but the smallest in sales contribution. Many items in this group registered extremely low or even negligible sales—some generating less than $5 in total sales during the entire year.

The segmentation revealed a highly concentrated sales distribution, where a small group of products (Category A) drove most of the revenue. This insight emphasizes the importance of carefully managing these best-sellers through effective inventory control, marketing efforts, and continuous availability. Conversely, the long tail of underperforming products (Category D) presents an opportunity for rationalization: these items could be bundled, discounted, or potentially removed from the assortment to reduce inventory complexity and free up resources.

Predictive Modeling

The predictive modeling stage aimed to forecast product sales volume using historical sales and financial data. Three algorithms were tested:

- **Linear Regression:** A simple, interpretable baseline.
- **Support Vector Regression (SVR):** Explored for capturing non-linear trends.
- **Random Forest Regressor:** A robust model suited for complex patterns and outliers.

Data Preparation for Modeling

Unlike the exploratory analysis, which focused on a single month, the modeling used data from all months between April 2024 and March 2025. This involved applying the same data wrangling processes across the entire period and aggregating all records into a single modeling dataset.

The prepared dataset consolidated sales history for each product, enabling the models to predict April 2025 sales based on patterns observed over the previous twelve months.

Features Used and Results

Key features included lagged sales (previous month, 12-month rolling averages), unit cost, unit profit, margin %, and time variables (cyclical month encoding and YearMonth). Product codes were target-encoded for model compatibility.

Predictive Modeling

After training and testing, Random Forest Regressor achieved the highest accuracy ($R^2 \approx 0.87$), outperforming the other models. This confirmed that sales variations could be explained through past sales patterns and product-level financial metrics.

Key Insight

This modeling step proved the feasibility of sales forecasting in retail. Future improvements could integrate seasonality adjustments, external factors like promotions, or more advanced forecasting models to further refine predictions.

Interactive Visualization and Reporting

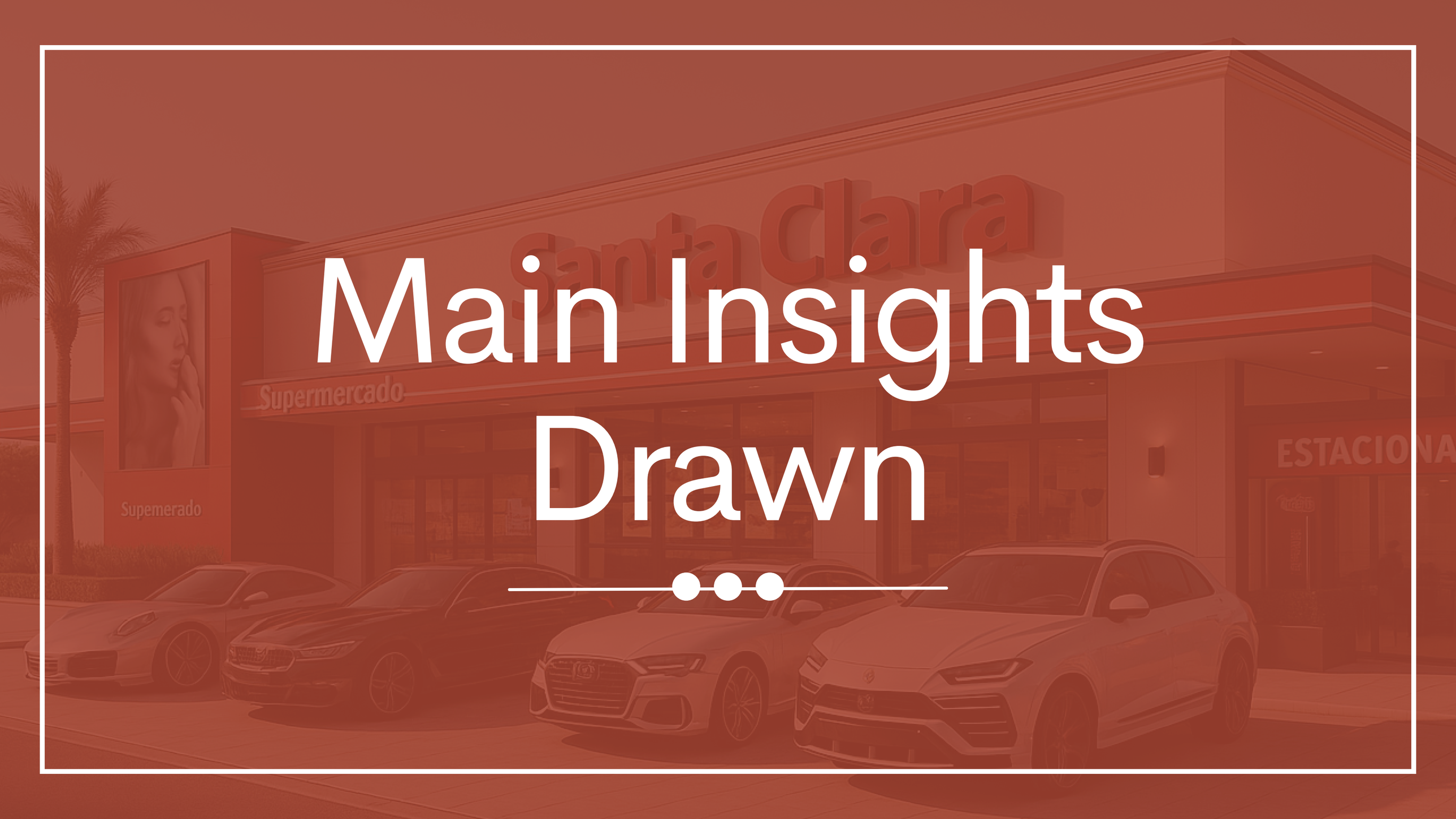
To make the analysis accessible and actionable, an interactive dashboard was built using Streamlit, allowing users to explore sales and profit performance dynamically. This tool enables both technical and non-technical stakeholders to analyze the supermarket's historical performance without writing code.

The dashboard offers the following features:

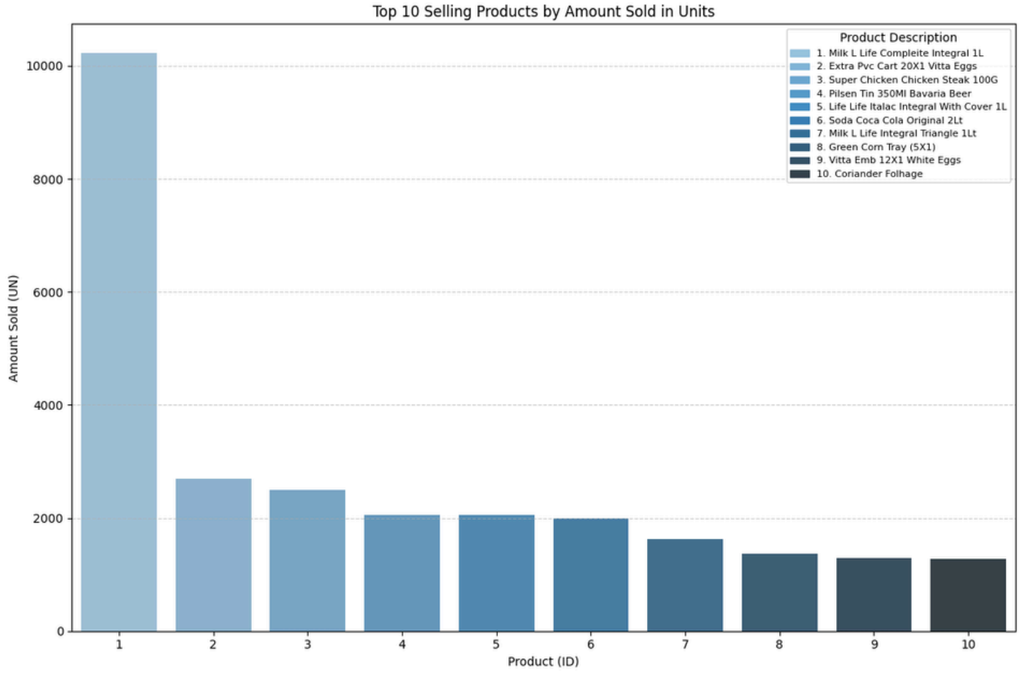
- Product-Level Analysis: Users can filter data by product code or description and select custom date ranges. This enables detailed exploration of specific products' sales, profit, and margin trends.
- Time-Series Visualizations: Interactive line charts display monthly trends in Total Sales, Total Profit, and Units Sold, helping users spot seasonality and performance fluctuations.
- Profitability Distribution: A combined boxplot and swarmplot visualize the monthly distribution of Margin %, highlighting outliers and variations in profitability.
- Predictive Insights: The dashboard integrates the linear regression model, offering a simple profit forecast and exposing the model's evaluation metrics (Mean Absolute Error, R^2) directly in the interface.
- Downloadable Data: Users can also explore the raw filtered data, promoting transparency and supporting deeper ad-hoc analysis.

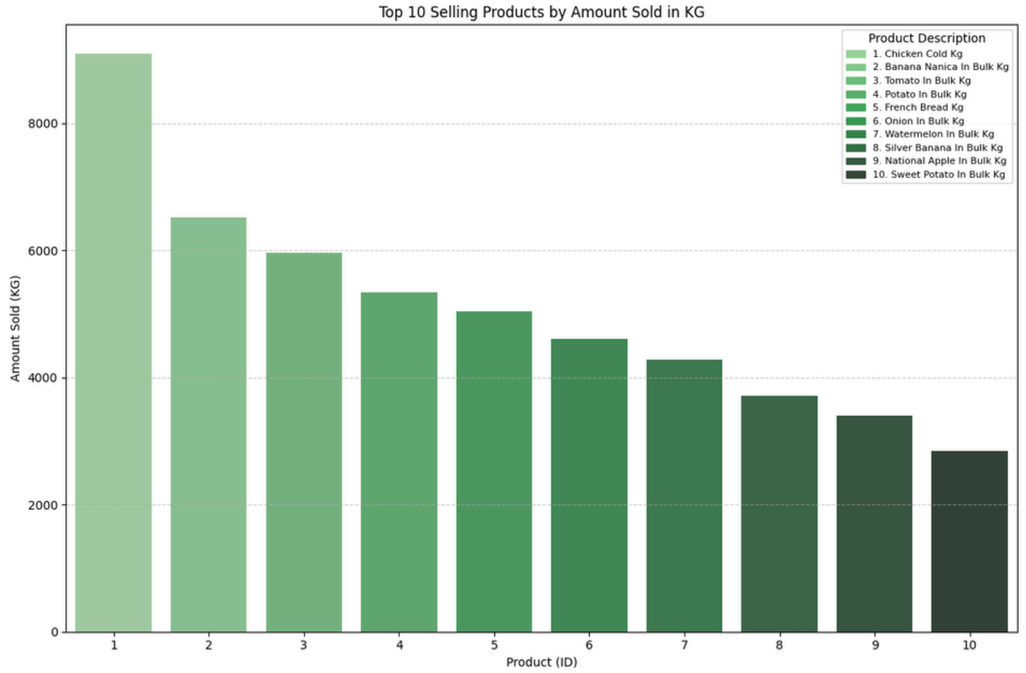
This interactive reporting tool transforms static analysis into a dynamic exploration experience, making it easier for decision-makers to extract insights and monitor business performance in near real-time.

Main Insights Drawn



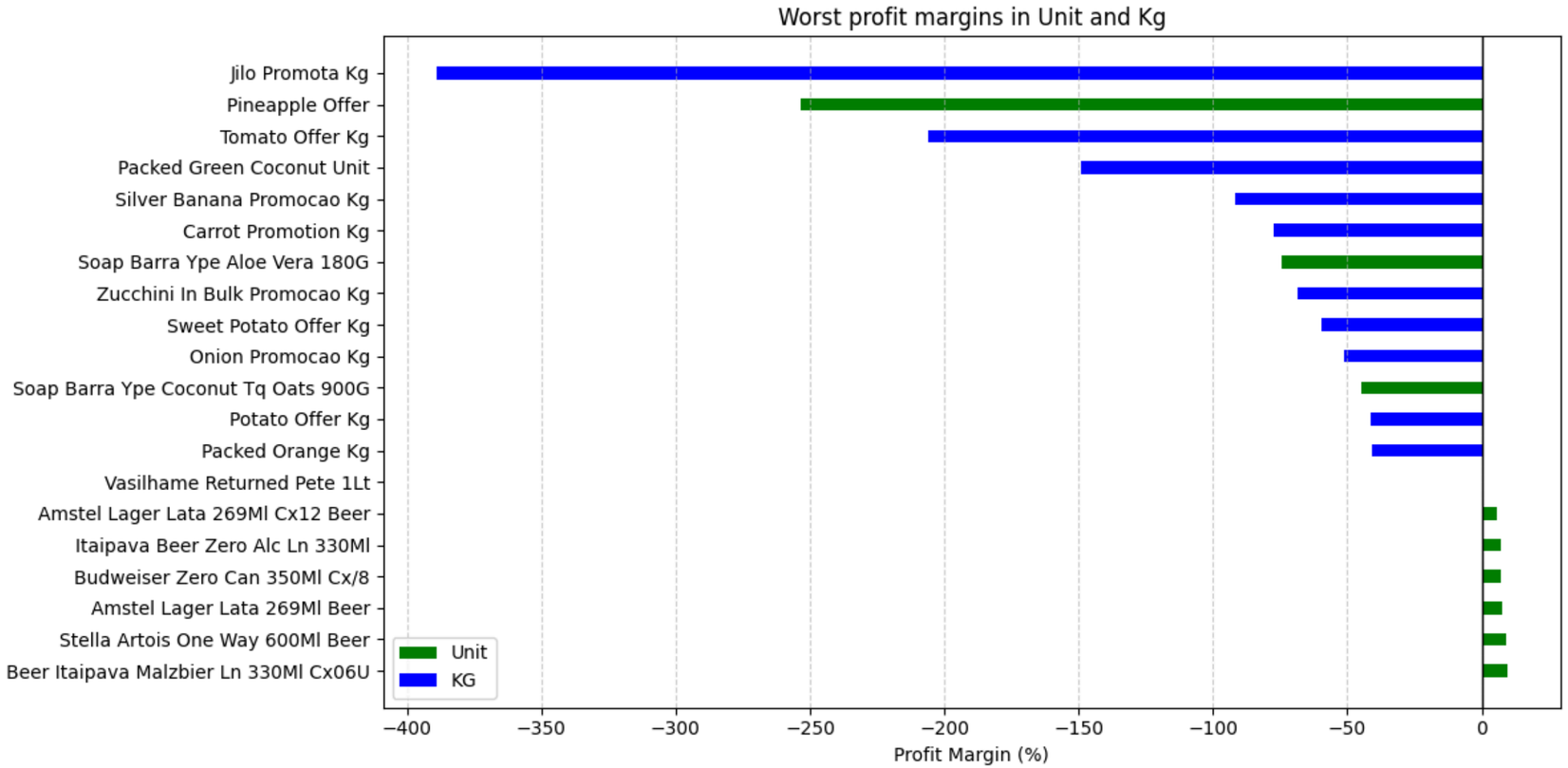
Top Selling Products by Amount Sold (Units & KG)





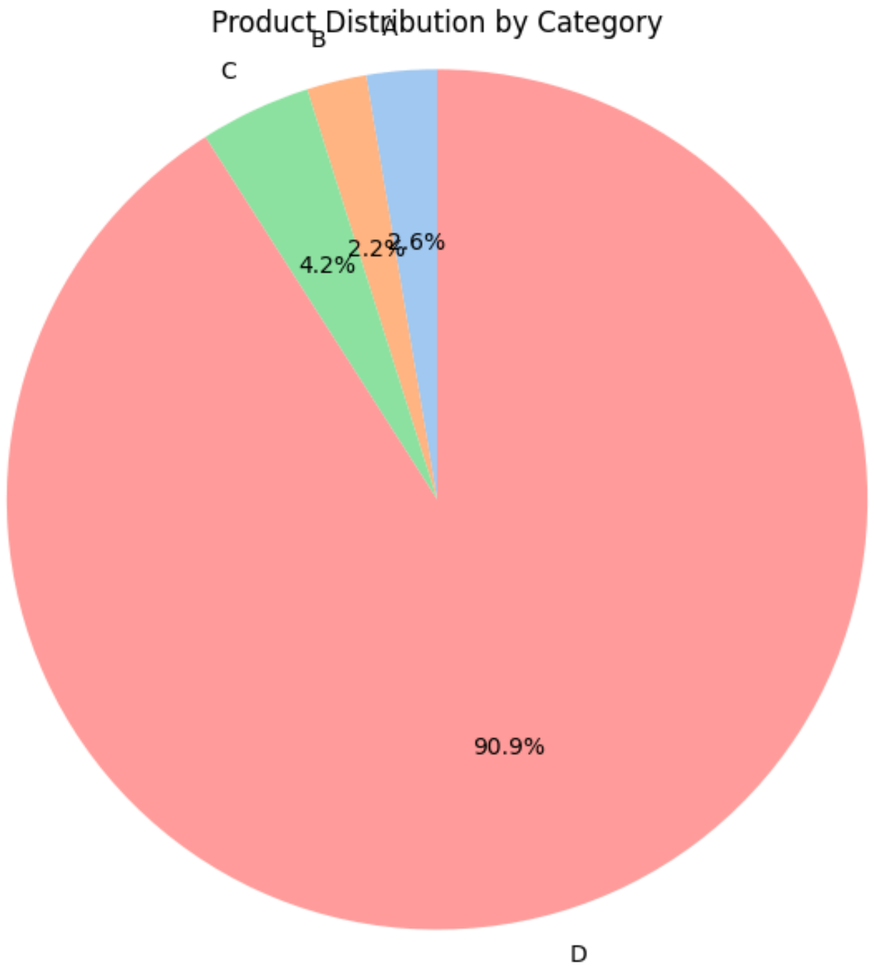
The most popular products, split between unit-based and weight-based sales, highlight core supermarket inventory drivers during the selected period.

Worst Profit Margins in Unit and Kg



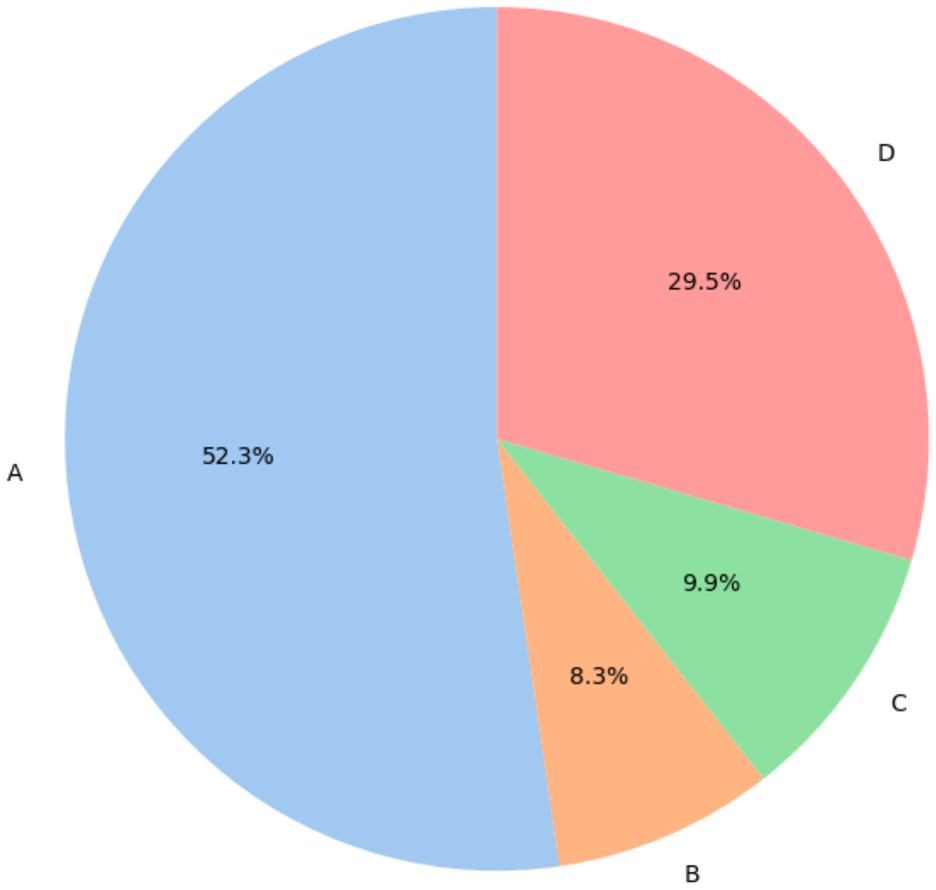
Products with the worst profit margins in March 2025. Some items, such as promotions and offers, sold well below cost, reflecting temporary discounts or possible pricing errors. Addressing these items is key to improving profitability.

Product Distribuition by Category



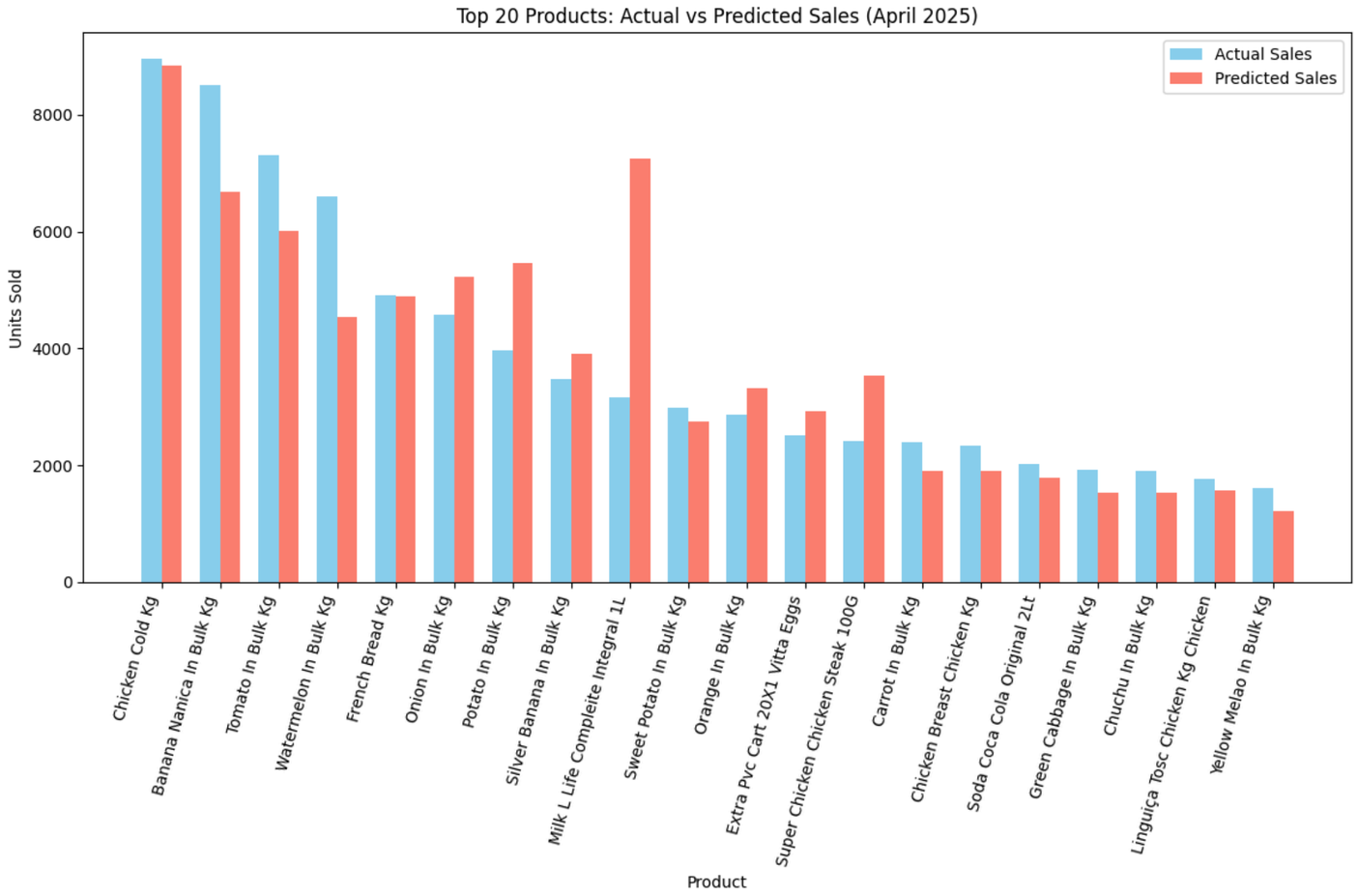
- This product segmentation reflects data from March 2025 only.
- The category distribution (A, B, C, D) may vary in other months depending on sales patterns, promotions, and product availability.
- Category A contains top sellers for this period, while Category D represents the long tail of low-performing items.
- For a complete view, this segmentation should be recalculated regularly to capture changes in product performance across time.

Sales Contribution by Category



- This chart shows each category's share of total sales revenue for March 2025.
- Even though Category A contains very few products, it generated over half of total sales (52.3%), confirming its strategic importance.
- Category D, while the largest in product count, contributed only 29.5% of revenue, reflecting the low sales performance of most products in this group.
- These proportions may shift in different months depending on demand, stock levels, and promotional activities.

Top 20 Products Actual vs Predicted April 2025

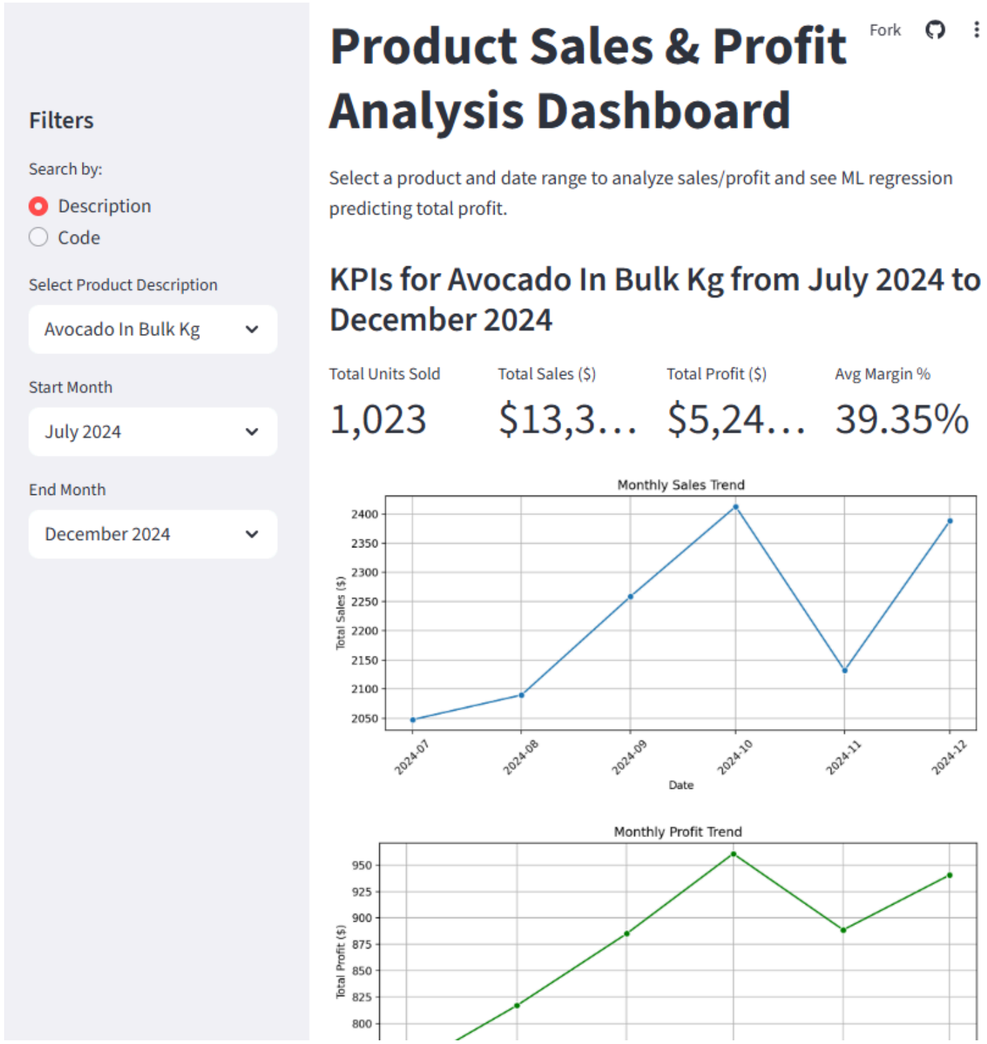


Comparison of actual and predicted sales for the top 20 products in April 2025.

The model achieved an R^2 score of 0.87, accurately capturing product-level sales patterns, with some deviations on specific items.

Confirms the model's potential for practical sales forecasting.

Interactive Dashboard Example



- In this example, the user selected "Avocado In Bulk Kg" for the period from July 2024 to December 2024.
- The dashboard displays key performance indicators (KPIs) including total units sold, total sales, total profit, and average margin %.
- Below, interactive time-series charts show monthly trends in sales and profit, enabling users to quickly assess product performance over the selected period.
- Filters on the left allow flexible analysis by product and date range, supporting dynamic and user-friendly exploration.

Conclusion

This project demonstrated how supermarket sales data can be turned into actionable insights through a complete data science pipeline. Using transactional data from April 2024 to April 2025, we uncovered product-level sales and profitability patterns, developed a predictive model, and built an interactive dashboard for easy analysis.

The exploratory analysis of March 2025 highlighted both top-selling products and those with the worst profit margins, showing that high sales volume does not always mean high profitability. A segmentation framework grouped products into four performance tiers, helping prioritize where to focus marketing, inventory, and pricing strategies.

On the predictive side, machine learning models were tested, with Random Forest achieving the best performance ($R^2 \approx 0.87$), accurately forecasting product sales volumes. The Streamlit dashboard allowed for flexible exploration of sales, profit, and model predictions, making these insights accessible to business users. Together, these components support data-driven decision-making in retail, offering a repeatable approach to understanding product performance, improving profitability, and forecasting future sales.

Thank you for your time