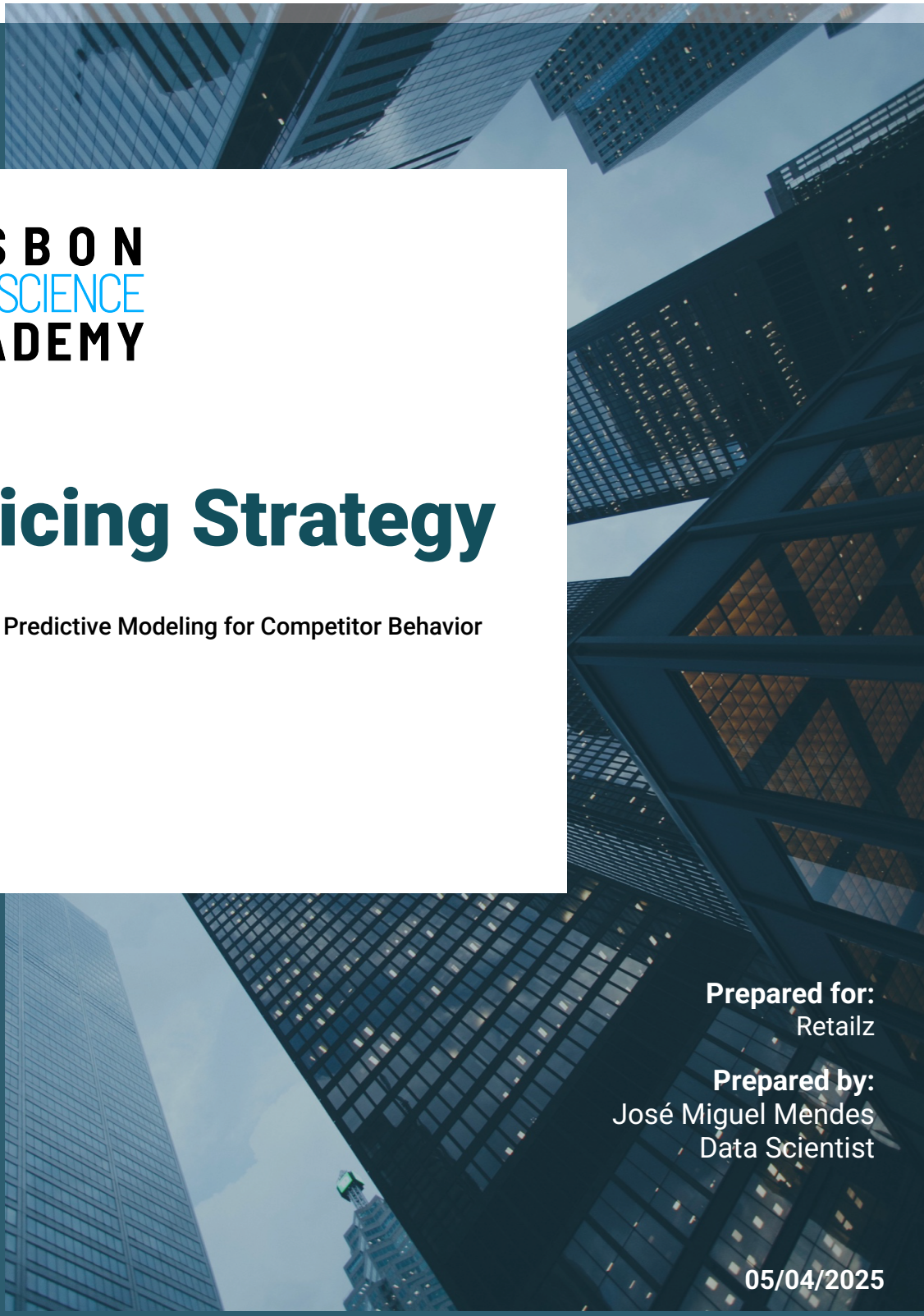

LISBON DATA SCIENCE ACADEMY

Retail Pricing Strategy

Campaign Effectiveness and Predictive Modeling for Competitor Behavior

Prepared for:
Retailz

Prepared by:
José Miguel Mendes
Data Scientist

05/04/2025

Table Of Contents

1. Client requirements

1.1 Summary

Retailz seeks to improve its pricing strategy through the historical analysis and forecasting of competitor prices. The project has two main objectives:

- Historical Analysis:
  - Understand competitor pricing behavior over time.
  - Identify discount patterns and promotional campaign dynamics.
  - Evaluate competitiveness variation across product categories.
  - Detect potential price reaction behavior among competitors (e.g., whether one competitor changes prices after another).
  - Compare average price levels of key competitors relative to Retailz across categories.
- Forecasting Solution via API:
  - Develop a system to forecast the selling price (pvp final) of two key competitors for a user-defined future date.
  - Forecasts should apply across multiple product categories with stable performance.
  - The model will be accessible via an API for easy integration with Retailz internal systems.

These objectives aim to provide analytical and predictive support to Retailz's pricing team, enabling them to anticipate market shifts and make proactive pricing decisions.

1.2 Requirements clarifications

To translate the business requirements into technical metrics and reduce ambiguity, we propose the following clarifications:

Requirement	**Technical Interpretation**	**Status**
Forecast by category	We will train models at SKU (Stock Keeping Unit) level, ensuring granularity and precision. structure_level_4 will be used only for analysis and aggregations across product families.	Confirmed
Forecast per competitor	The model will predict individual prices for CompetitorA and CompetitorB by SKU and date.	Confirmed
Seasonality & reactive behavior	Statistical tests and time-series analysis will identify price reaction dynamics.	Confirmed
Performance across categories	Forecast performance MAE (Mean Absolute Error) or RMSE (Root Mean Square Error) will be evaluated per category.	Confirmed
Campaign data availability	Available only for Chain and CompetitorA. CompetitorB will be modeled without campaign features.	With limitation
Quantity measurement	Due to lack of unit definitions, quantity will be used only for aggregated analysis.	Limited use
API delivery	Forecasts will be served through a REST endpoint with SKU/category and date as input.	Planned

2. Dataset analysis

2.1 General analysis

Dataset Description

The sales dataset contains product sales data spanning the period from January 3rd, 2023 to October 28th, 2024. The data has a daily granularity, with each row representing the quantity sold (quantity) of a specific SKU on a given day.

- Total distinct SKUs: 3,605.
- Date range: 664 days.
- Days with quantity = 0**:** only 12/25/2023 and 01/01/2024.
- Rows removed due to invalid values (quantity < 0): 292.

Data Cleaning

- Negative quantity values were treated as input errors and removed.
- Days with no sales were treated as expected operational gaps, not as missing data.

Product Hierarchy

Products are organized in a hierarchical category structure. The most specific level, structure_level_4, was used in the analysis as it groups similar products, allowing for more consistent aggregation and trend interpretation.

Chart 1: Top 5 chain Structure Level 4 with Highest Median Daily Growth (Monthly)

Preliminary Observations

- Sales fluctuate as expected around holidays.
- Some product families (structure_level_4) concentrate a higher volume of sales.
- General upward trend over time, with variations between categories.

Campaign Analysis

In addition to sales data, promotional campaign datasets were analyzed for Chain and Competitor A**.**

- Average discount per campaign**:**
  - Chain - C2: 27.7%, C1: 27.3%.
  - CompetitorA - A1: 26.9%, A2: 26.7%, A3: 26.3%.

- Median and most frequent campaign duration**:**
  - Chain (C1, C2)**:** both median and mode = 6 days.
  - Competitor A**:**
    - A1 and A2: median and mode = 2 days.
    - A3: median = 2.5 days, mode = 3 days.

*Chart 2**:** Median Campaign Duration by Campaign Type*

Chart 3: Most Frequent Campaign Duration by Campaign Type

- Campaign start day patterns:
  - Competitor A: Fridays dominate (n = 39), followed by Thursdays (3), Saturdays (2), and one on Tuesday.
  - Chain: campaigns typically start on Tuesdays (n = 30) and Mondays (n = 24).

Chart 4: Distribution of Start Day by Campaign Type

- Temporal gaps and seasonality:
  - Competitor A:
    - Had no campaigns starting in April, June, or July of 2024.
    - March 2023 had the highest campaign count overall (n = 5).
    - In 2024, January had the highest count (n = 4).
    - A noticeable decline in campaign frequency was observed throughout 2024.
    - In March 2024, campaign A2 remained active for the entire month - an outlier in duration.

Chart 5: Distribution of Campaigns by Month and Competitor

2.2 Business questions analysis

Based on your questions, here are our answers backed by data:

Which competitors tend to be more expensive/cheaper?
On average, Competitor A has higher final prices than Chain, while Competitor B tends to be cheaper.

Looking deeper into product categories (up to structure level 2), the analysis shows a consistent pattern where Competitor B tends to be the cheapest across almost all categories, while Competitor A and Chain alternate as the most expensive depending on the category. For example, in category 101, Competitor B offers the lowest prices, while Chain is the most expensive. In categories 102 and 103, Competitor B remains the cheapest, but Competitor A takes the lead as the most expensive. This suggests that Competitor B generally competes aggressively on price across product categories, whereas Competitor A and Chain position themselves differently depending on the category.

Chart 6: Average Final Price Difference (CompetitorA - Chain) by Month

Chart 7: Average Final Price Difference (CompetitorB - Chain) by Month

Image 1: Competitor Price Comparison per Category (Structure Level 2)

What is the average discount per competitor across product categories?

Analysis of average discounts reveals distinct promotional strategies among competitors across product categories up to structure level 2. Competitor A tends to offer higher average discounts in many categories, reflecting a more aggressive pricing strategy to attract customers. Chain shows relatively moderate discount levels, while Competitor B generally applies smaller discounts, indicating a more conservative approach.

Image 2: Discount Analysis by Competitor and Category (Structure Level 2)

How does price vary depending on promotional strategies?

Hypothesis tests reveal significant differences in prices between promotional and non-promotional periods across competitors and product categories (structure level 2). In most cases, promotional periods are associated with significantly lower median prices, confirming the effectiveness of promotions in reducing consumer costs.

However, a few exceptions were observed where the median promotional price was not lower, suggesting the use of alternative promotional strategies or different types of offers. These findings highlight that price adjustments during promotional events are a common and impactful tactic among competitors.

Chart 8: Median Price Comparison: Promotion vs No Promotion by Category and Competitor

Is there reactive behavior among competitors?
Qualitative patterns suggest cyclical and seasonal responses, especially between Chain and Competitor A. Campaigns often occur in alternating sequences, hinting at strategic reactions. For example, Chain campaigns tend to start on Mondays or Tuesdays, whereas Competitor A predominantly initiates campaigns on Fridays, possibly as a counter-strategy.

A full causal attribution would require time-series intervention models to validate these reactive dynamics.

What is the impact of campaigns on sales?
Statistical analysis using hypothesis testing and regression showed that:

- Competitor A's campaigns have an even stronger positive spillover effect.
- Simultaneous campaigns amplify demand (synergistic effect).

Chart 9 : Hypothesis test – Chain's campaign inactive

Chart 10: Chain's campaign inactive – Comparison of Median Sales by Competitor's Campaign Status

Chart 11 : Hypothesis test – Chain's campaign active

Chart 12: Chain's campaign active – Comparison of Median Sales by Competitor's Campaign Status

Is there seasonality in promotions?
Yes. Promotion cycles are particularly evident for Competitor A, with a ~77-day recurrence between campaigns. Competitor A's campaigns are less frequent in 2024, suggesting a shift in strategy or budget constraints.

- March 2023 and January 2024 had the highest campaign volumes.
- March 2024 saw campaign A2 active throughout the entire month**,** a unique deviation from typical durations.
- Chain maintains a more regular campaign rhythm, typically initiating on weekdays.

Chart 13: Time Series Decomposition (CompetitorA Campaigns)

Chart 4: Distribution of Start Day by Campaign Type

Chart 5: Distribution of Campaigns by Month and Competitor

2.3 Recommendations

Based on our analysis, we recommend the following:

- **Prioritize high-CDGR categories**

Focus marketing and pricing strategies on product categories (structure_level_4) with the highest Compound Daily Growth Rate (CDGR), as they show organic momentum and responsiveness to promotional activity.

- **Improve competitor monitoring and timing**

Competitor A's campaigns exhibit a clear cyclical pattern (~77-day interval) and typically begin on Fridays. Automated tracking of their behavior would enable more agile responses. We highlight a notable gap in historical campaign activity from April to July 2024, which may represent a window of opportunity for Chain to gain share. However, statistical tests indicated that Chain's sales during these inactive periods were not consistently or significantly higher, suggesting that timing alone may not guarantee uplift — strategic action is still required.

- **Optimize campaign scheduling**

Chain's own campaigns are typically launched on Mondays and Tuesdays and last 6 days. We recommend testing adjusted timing strategies, such as:

- Launching campaigns slightly ahead of known Competitor A patterns.
- Experimenting with shorter, high-intensity promotions to match Competitor A's urgency.
- Avoiding saturation by spacing campaigns across underused weekdays (e.g., Wednesday).

- **Strategically leverage simultaneous promotions**

Data suggests that simultaneous campaigns amplify demand, rather than canceling each other out. Consider coinciding promotions in peak-demand periods or in high-traffic product categories to maximize volume uplift.

- **Evaluate and refine campaign ROI**

While not all Chain campaigns show statistically significant uplift, many have positive directional effects. We recommend:

- Performing category-level ROI analysis, especially in high-CDGR segments.
- Evaluating marginal return per discount unit to avoid over-discounting.
- Testing differentiated strategies (e.g., bundling vs. pure discounts).

3. Modelling proposal

3.1 Modelling task

The primary modelling task is to forecast the final selling price (pvp_final) of competitor products for a given SKU and future date, specifically for CompetitorA and CompetitorB. This forecast will power an API that allows Retailz to query expected prices on a user-defined target date, across multiple product categories.

This enables Retailz to:

- Proactively adjust its pricing to remain competitive.
- Identify pricing trends and anticipate competitive behavior.
- Inform promotional decisions based on predicted future price movements.

To meet this need, we will develop forecasting models at SKU level, ensuring high granularity and allowing for aggregation into structure-level insights if required. The solution will be designed to perform consistently across product families, satisfying the requirement for reliable price predictions across categories.

3.2 Modelling strategy

Problem framing

The task is a regression-based time series forecasting problem at the SKU level, where the target variable is pvp_final for both CompetitorA and CompetitorB. This is framed as a supervised learning problem, leveraging historical pricing behavior and temporal dynamics to support predictive modeling.Approach and candidate models

Suggested Modeling Approaches:
Given the characteristics of the data and the forecasting objective, the following modeling families are proposed for consideration:

- Classical time series models (per SKU):
    - ARIMA/SARIMA **-** May be considered for illustrative or benchmarking purposes. However, due to the limited historical window (22 months), their applicability may be restricted, especially in modeling seasonality.
- Random Forests**:** Recommended as a baseline approach due to their robustness and interpretability.
- Machine learning models
    - *Gradient Boosted Trees (XGBoost)* – Suitable for capturing complex, non-linear patterns in pricing behavior across SKUs..

Although a general modeling strategy is initially proposed, the possibility of segmenting SKUs based on data characteristics (such as history length or variability) may be explored if it leads to significant performance gains.

Feature Engineering

We anticipate generating the following featuresLag features**:**

- Previous pvp_final values (e.g., 1-day, 7-day, 14-day lags).
- Rolling statistics**:** Rolling mean, std, min, max of price over recent windows.
- Calendar features**:** Day of week, week of year, month, holidays.
- Campaign indicators**:** Binary flags for active campaigns from Chain or competitors.
- Seasonality indicators**:** Flags for known cyclical patterns (e.g., CompetitorA's 77-day cycle).

Data Preparation and Imputation Strategy

Before modeling, it is important to ensure that the training dataset contains complete price information (pvp_final) for all three entities — the Chain, CompetitorA, and CompetitorB — across all SKU and date combinations. While Chain prices are generally complete, preliminary data exploration suggests substantial gaps in competitor price records.

To address this, we propose using a supervised machine learning imputation model, with Random Forest as a candidate method. The model would be trained on historical records where competitor prices are available, using the following features:

- Price from the Chain for the same SKU and date.
- Temporal variables (e.g., holidays, weekdays, month).
- Campaign indicators (e.g., is_promo, campaign type).
- Indicators of seasonal behavior.

This approach is expected to predict missing prices with high accuracy, resulting in a complete dataset where each row represents a (time_key, SKU, competitor) combination. The final dataset would include:

- Actual prices for the Chain.
- Actual or imputed prices for CompetitorA and CompetitorB.
- A flag indicating whether a competitor price is real or imputed.

If validated through metrics such as MAE and R^2, the selected imputation approach would serve as the standard method for completing competitor price data and would provide the input data foundation for subsequent forecasting models.

Model evaluation and selection

Model performance will be evaluated at the SKU level using both absolute and relative error metrics, with a focus on ensuring numerical accuracy in predicted prices:

- **Mean Absolute Error (MAE)** will be used as the primary metric, as it measures the average absolute difference between predicted and actual prices in the same units. This makes it directly interpretable and relevant from a business perspective, especially when evaluating pricing precision at the individual SKU level.

- **Symmetric Mean Absolute Percentage Error (sMAPE)** will be employed as a complementary metric to capture relative prediction accuracy, which is useful when comparing across SKUs with different price levels.
- **Root Mean Squared Error (RMSE)** will be considered to highlight larger errors, though it will be less emphasized in final model selection.

Evaluation will be performed using time-series cross-validation with forward splits on historical data, ensuring the model is tested on past periods that were not seen during training while maintaining temporal consistency.

Model selection will be based on:

- Consistent reduction in MAE across SKUs.
- Minimal increase in relative error (sMAPE), even when training data is limited.
- Overall enhancement in point prediction accuracy, supporting models suitable for production deployment.

Expected Model Performance

Based on preliminary analyses and domain knowledge, we anticipate the forecasting models to achieve a level of accuracy suitable for supporting operational decision-making in competitive pricing. Specifically:

- **Accuracy:** We expect the models to produce Mean Absolute Errors (MAE) in the range of 3 to 5 price units per SKU, which aligns with observed historical price variability and business tolerance for prediction error. This level of precision should allow Retailz to make informed pricing adjustments with reasonable confidence.
- **Relative error:** Complementary metrics such as sMAPE are expected to remain below 12%, indicating that relative prediction errors across SKUs with diverse price points will be within acceptable bounds.
- **Robustness:** The models should maintain stable performance across SKUs and temporal periods, including promotional events and pricing volatility, thanks to the inclusion of calendar and campaign features.
- **Adaptability:** Given the periodic retraining strategy, the models will adapt over time to changes in competitor pricing policies and market dynamics, limiting degradation in performance due to strategic shifts.

While some variability in accuracy is expected across SKUs with limited historical data or highly irregular pricing patterns, the combination of granular SKU-level modeling and comprehensive feature engineering aims to deliver consistent and actionable forecasts.

Future validation during deployment phases will provide further insight into real-world performance and guide refinements to maintain and enhance accuracy.

3.3 Expected issues and risks

Several risks have been identified that may affect model development:

- Implicit seasonality: not all seasonal events are labeled — risk of underfitting seasonal patterns.
- Missing data: even with imputation, bias may arise if missing prices are systematically different.
- Price reactivity: if strong interdependence exists among competitors, independent models may underperform.
- Future strategy shifts**:** model may be sensitive to abrupt changes in competitor pricing policies.

Mitigation via modelling and feature engineering strategy

The pipeline was designed to address these challenges as follows:

- Implicit seasonality: The feature engineering process includes time-based indicators (month, dayofweek) and custom flags for key events (is_christmas_season, is_black_friday, etc.) to capture recurring seasonal effects even when not explicitly labeled.
- Missing data**:** Missing values are handled via a preprocessing pipeline that applies mean imputation, ensuring robustness while maintaining simplicity and interpretability.
- Price reactivity**:** The model incorporates lag-based features (lag_diffB, lag_diffB_sl4) and competitor pricing (chain_price) to reflect historical interactions and price dynamics between products and competitors.
- Strategy shifts**:** While the pipeline is not yet automated, the model will be periodically retrained with updated data to reflect changes in market behavior or competitor pricing strategies.

Annexes

Chart 1: Top 5 chain Structure Level 4 with Highest Median Daily Growth (Monthly)

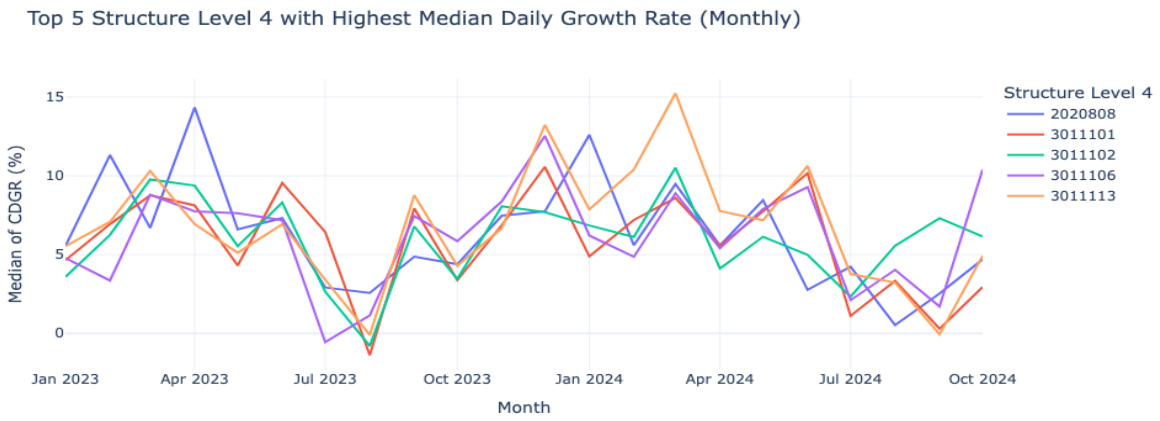

Chart 2: Median Campaign Duration by Campaign Type

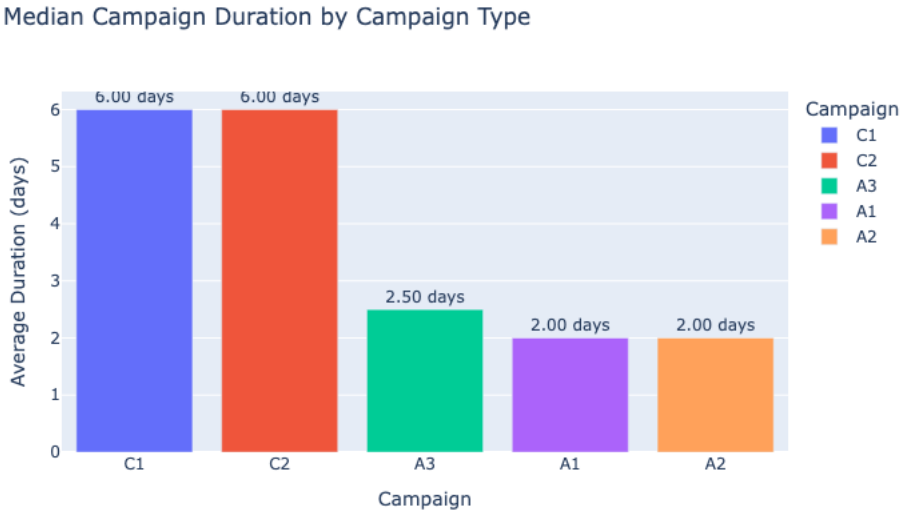

Chart 3: Most Frequent Campaign Duration by Campaign Type

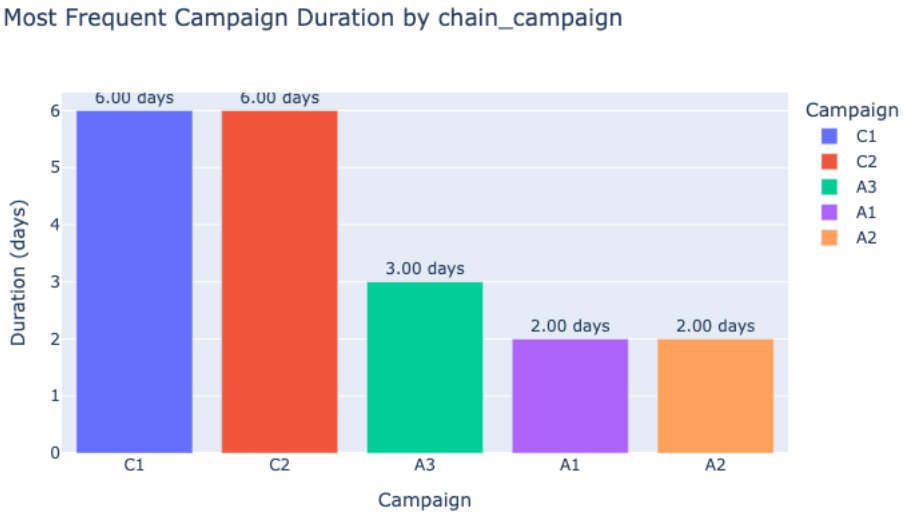

Chart 4: Distribution of Start Day by Campaign Type

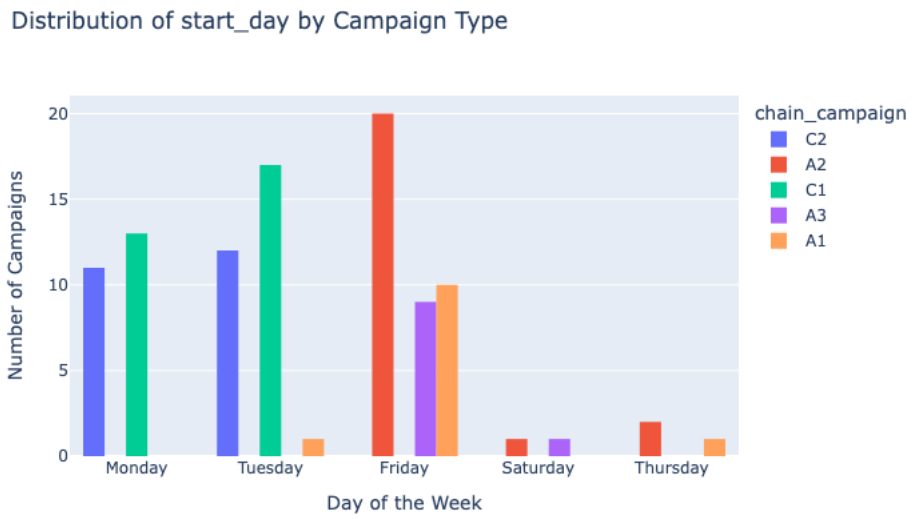

Chart 5: Distribution of Campaigns by Month and Competitor

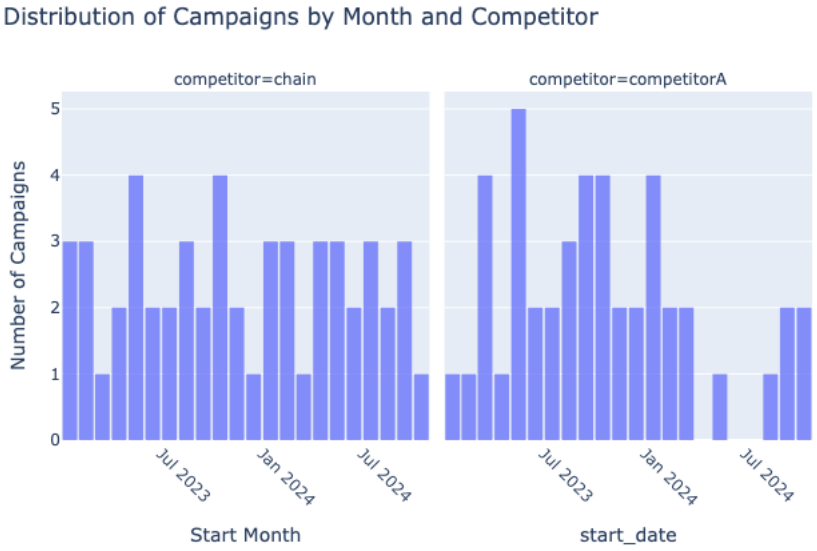

Chart 6: Average Final Price Difference (CompetitorA - Chain) by Month

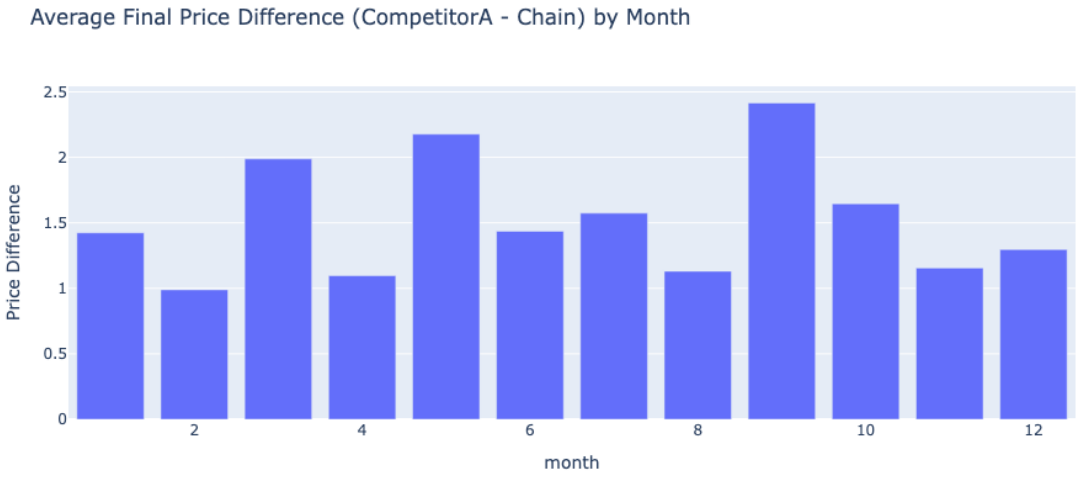

Chart 7: Average Final Price Difference (CompetitorB - Chain) by Month

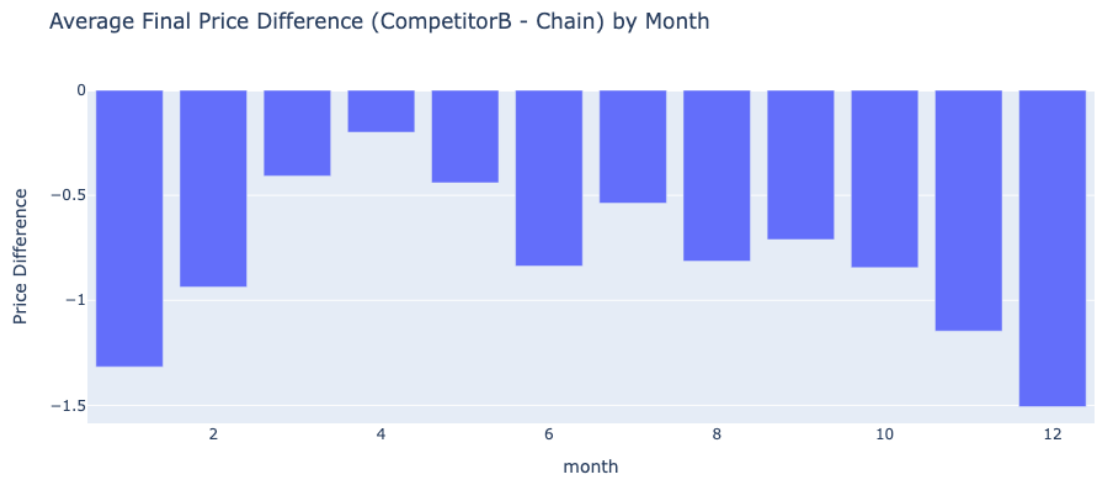

Image 1: Competitor Price Comparison per Category (Structure Level 2)

structure_level_2	cheapest_competitor	most_expensive_competitor
101.0	competitorB	chain
102.0	competitorB	competitorA
103.0	competitorB	competitorA
104.0	competitorB	competitorA
105.0	competitorB	chain
106.0	competitorB	chain
201.0	competitorB	competitorA
202.0	competitorB	chain
301.0	competitorB	competitorA
302.0	competitorB	chain
303.0	competitorB	competitorA
304.0	competitorB	competitorA
305.0	competitorB	chain

Image 2: Discount Analysis by Competitor and Category (Structure Level 2)

competitor	chain	competitorA	competitorB
structure_level_2			
101.0	0.014329	0.042124	0.024602
102.0	0.007890	0.014301	0.018636
103.0	0.017523	0.020133	0.008602
104.0	0.002357	0.002475	0.014265
105.0	0.003644	0.003207	0.010836
106.0	0.006029	0.011241	0.003977
201.0	0.206327	0.046415	0.065585
202.0	0.085705	0.075879	0.025063
301.0	0.094702	0.051447	0.021387
302.0	0.101494	0.064541	0.014821
303.0	0.132129	0.067107	0.033093
304.0	0.122963	0.078550	0.013557
305.0	0.053453	0.029344	0.018556

Chart 8: Median Price Comparison: Promotion vs No Promotion by Category and Competitor

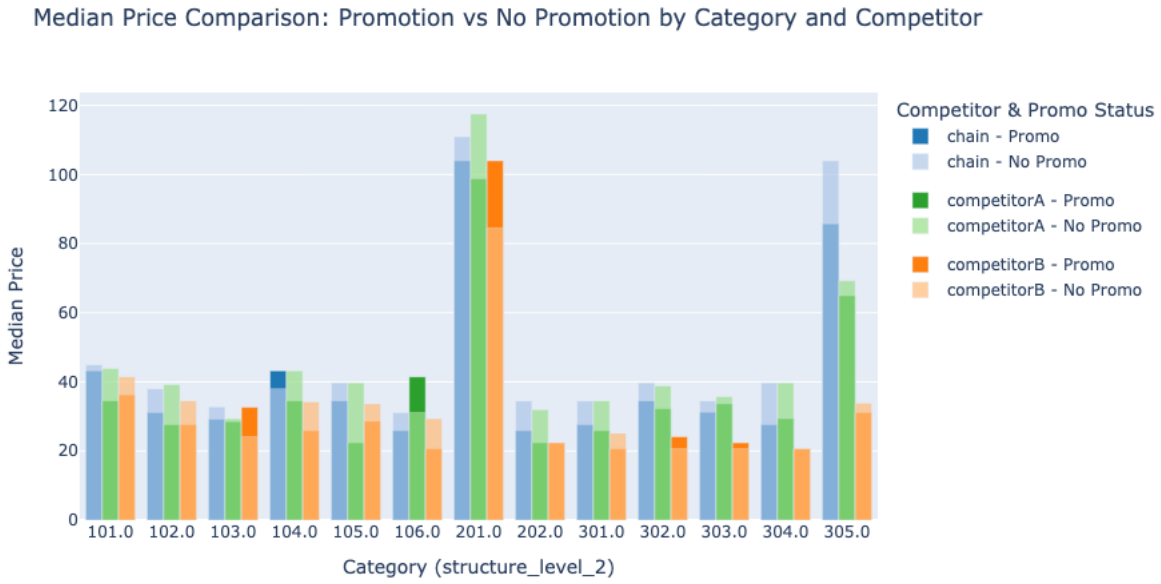

Chart 9 : Hypothesis test – Chain's campaign inactive.

Comment: Even when the chain's campaigns are inactive, the results show that competitor activity significantly influences sales in more than half of the analyzed segments (58% with H0 rejected).

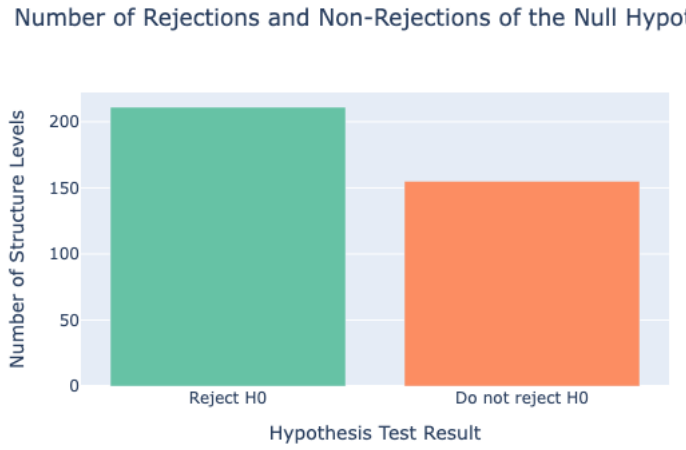

Chart 10: Chain's campaign inactive – Comparison of Median Sales by Competitor's Campaign Status.

Comment: Even more striking is that in 82% of cases, sales were higher when the competitor had an active campaign, suggesting a possible indirect positive effect or a spillover in demand.

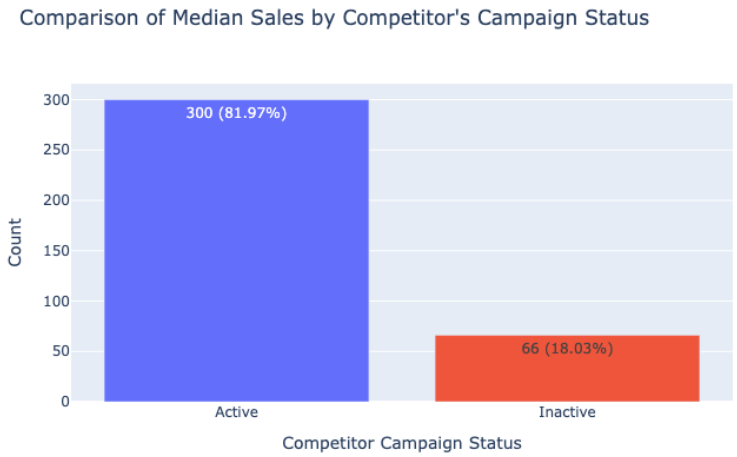

Chart 11 : Hypothesis test – Chain's campaign active.

Comment: When the chain has an active campaign, competitor activity still plays a relevant role: in 65% of segments, there was a statistically significant difference in sales.

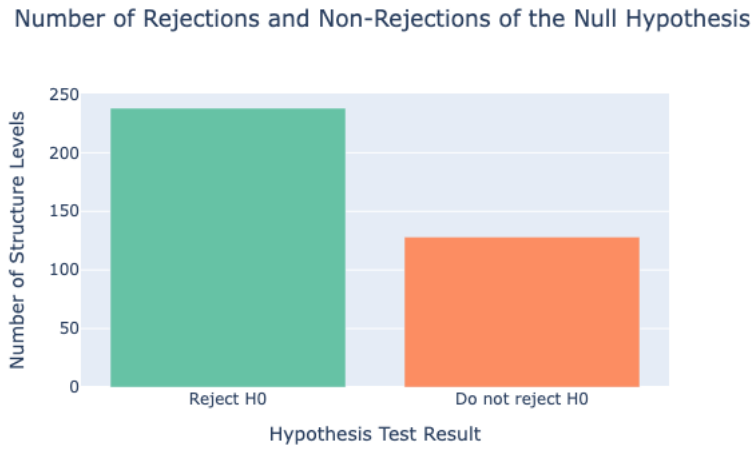

Chart 12: Chain's campaign active – Comparison of Median Sales by Competitor's Campaign Status.

Comment: In 83% of cases, the median sales were higher when the competitor was also active.

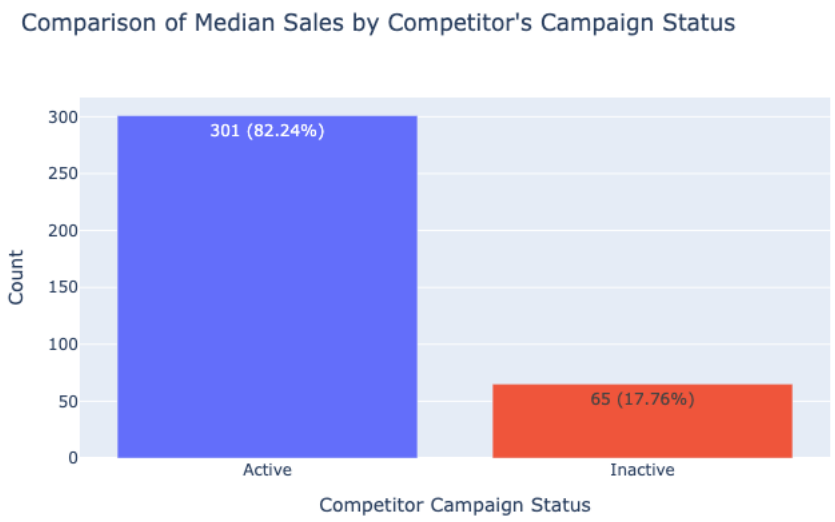

Chart 13: Time Series Decomposition (CompetitorA Campaigns)