# **Introduction**

In the food business, predicting sales may be challenging, particularly when juggling perishable inventory with erratic consumer demand. The biggest grocery operator in Ecuador, Corporación Favorita, wants to use machine learning to increase forecasting accuracy. Participants in this Kaggle challenge must forecast the daily unit sales of thousands of goods across many retail locations. Historical sales, promotions, holidays, oil prices, and other information are all included in the data collection. Reliable projections become crucial since the grading standards give perishable goods a higher priority. A competition like this provides a practical window of opportunity for retail operations to be streamlined and for data analytics to guarantee consumer happiness.

# **Pre Processing**

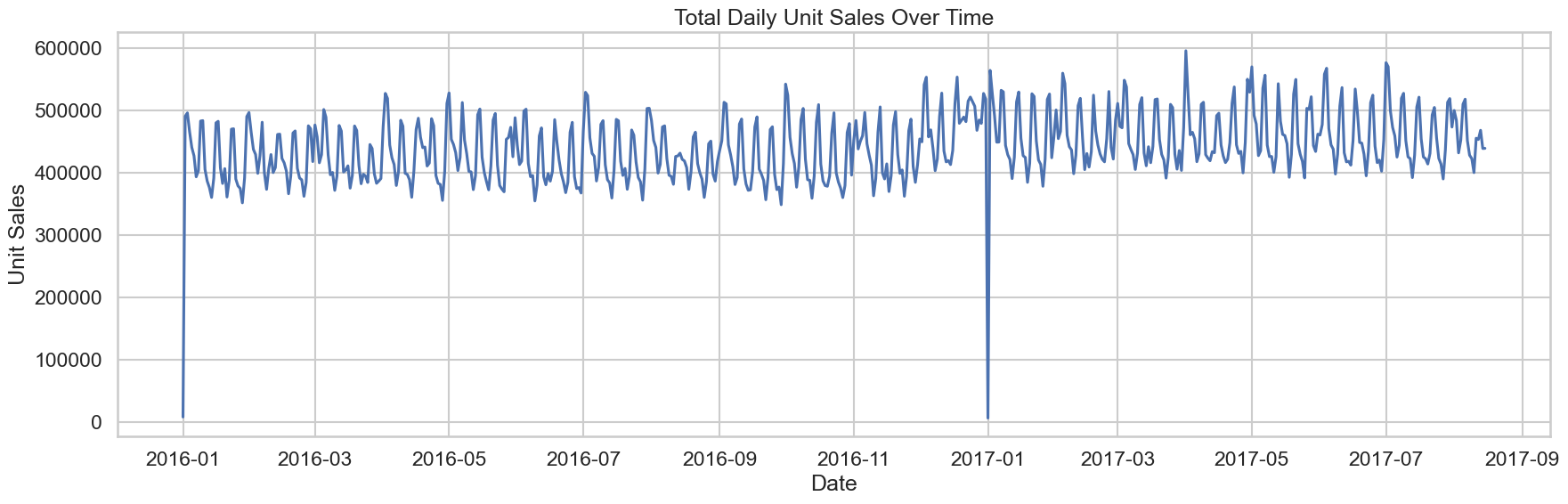
The cleaning process begins by handling missing values in the onpromotion column, which is filled with False and converted to a boolean type. This ensures consistency and avoids issues during modeling. Next, the focus shifts to the unit\_sales column, which contains the target variable for the forecasting task. Since the dataset includes both returns (negative sales) and extreme values, it's essential to clean these irregularities to build a more robust model.

Outliers are identified using the Interquartile Range (IQR) method, which is effective for detecting unusually high or low values. The IQR is computed as the difference between the 75th percentile (Q3) and 25th percentile (Q1). Values lying outside 1.5 times the IQR below Q1 or above Q3 are considered outliers. This method revealed over 11 million outliers, indicating significant sales variance likely due to seasonal spikes, promotions, or data entry errors.

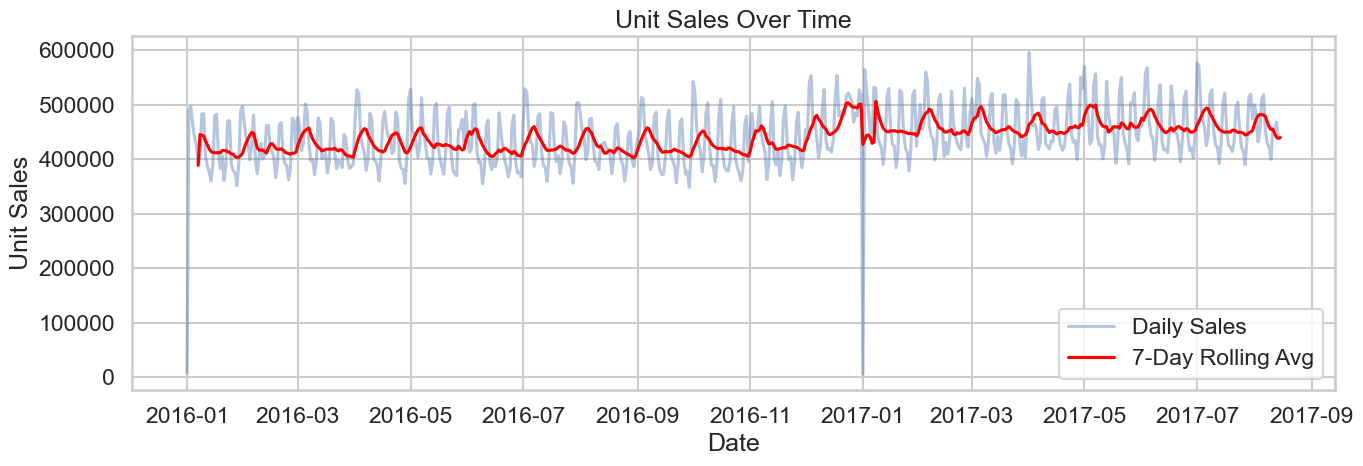
Additionally, 7,795 rows had negative unit sales, likely representing product returns. These are removed since the goal is to forecast actual sales rather than returns. The dataset is then filtered to include only non-negative, non-outlier sales data, resulting in a cleaner and more stable training set. This step is crucial to prevent skewed learning and improve model generalization.

The dataset is enriched by merging store and item metadata, adding contextual features like city, type, and perishable status. Irrelevant columns are dropped, and the data is filtered from 2016 onward. This reduces memory usage, ensures relevancy, and prepares a clean, feature-rich dataset for effective model training and analysis.

# **EDA**

***Temporal analysis***  


There is a noticeable weekly seasonality in the total daily unit sales, with peaks and troughs occurring every seven days. From January 2016 until about June 2017, the sales level tended to remain steady, fluctuating within a predictable range. Early in January 2017, there is a noticeable steep outlier, a strong decrease followed by a quick rebound, which represents a transient disruptive event. Although overall sales levels remained unchanged, there is a little indication that daily sales values in the second half of 2017 were more volatile than those in the first half of the observation period. This implies that over the later portion of the observation period, the daily variations may have grown substantially in amplitude. Sales dipped on the holiday (expected). Bounce back on 2017-01-02 shows pent-up demand — people probably shopped the day after. This is a clear holiday effect pattern, and could be a useful feature signal in modeling



The time series plot reveals a clear weekly seasonality, with peaks and troughs occurring roughly every seven days—suggesting higher sales on specific weekdays. Despite daily fluctuations, the overall trend from early 2016 to late 2017 appears relatively stable, showing no significant long-term upward or downward movement. The 7-day rolling average (red line) smooths out short-term noise, making consistent patterns more visible. A sharp drop in early January 2017 stands out as a potential outlier or anomaly, possibly tied to an external event. Additionally, there appears to be a slight increase in volatility during late 2017, though the rolling average remains steady.

A graph of a graph showing the average unit sales by day of week

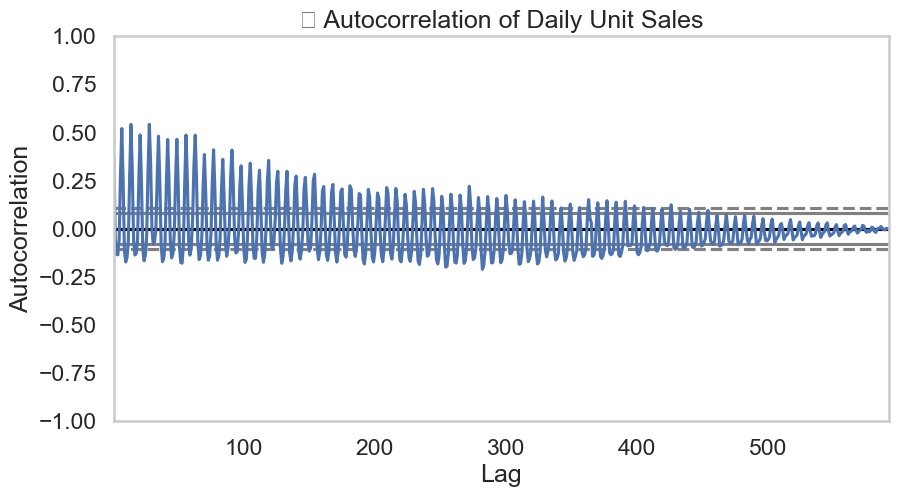
AI-generated content may be incorrect.

The bar chart reveals a distinct weekly sales pattern, with lower average unit sales from Monday to Thursday. Sales begin rising on Friday and peak over the weekend, especially on Sunday. This suggests increased consumer activity during weekends, highlighting the importance of incorporating day-of-week effects in sales forecasting models.

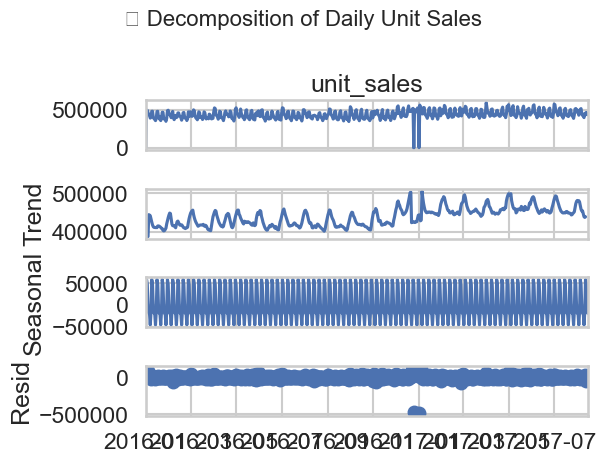
A graph with a line and a blue line

AI-generated content may be incorrect.

he monthly total unit sales show a stable yet mildly cyclical pattern. Peaks are often observed around December and mid-year (June or July), suggesting seasonal consumer behavior, possibly tied to holidays and mid-year events. The overall sales remain within a relatively consistent range, indicating stable market demand with periodic boosts. This recurring fluctuation highlights the importance of incorporating seasonal factors into forecasting models for improved accuracy.



The autocorrelation plot confirms strong weekly seasonality, with significant positive spikes at lags of 7, 14, 21, etc., indicating high correlation with sales from previous weeks. An oscillating pattern emerges, with alternating positive and negative correlations, typical of seasonal data. The gradual decay in correlation suggests diminishing influence over time. Many values exceed the confidence bounds, highlighting statistically significant autocorrelations and reinforcing the non-random, structured nature of the sales time series.

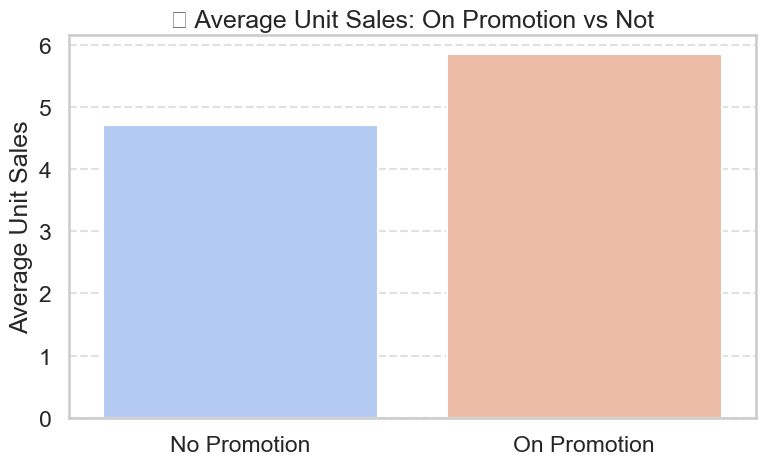


The decomposition of daily unit sales reveals a stable underlying trend with no significant long-term growth or decline. A strong weekly seasonality is evident, with consistent cyclical patterns repeating every 7 days. The residuals mostly resemble random noise, but notable spikes—especially around outlier periods—suggest external disruptions or special events not captured by trend or seasonality. These findings confirm that weekly patterns significantly influence sales behavior, and any forecasting model should incorporate this strong seasonality to improve accuracy.

ADF Statistic: -3.459269762331552

p-value: 0.009096861227454676

The ADF test (p = 0.009) confirms the daily sales data is stationary, making it suitable for time series modeling. A seasonality strength of 0.496 indicates moderate to strong weekly seasonality, explaining nearly half the data’s variability. These findings support the use of models that incorporate seasonality to effectively capture the structured, recurring patterns in sales behavior for improved forecasting accuracy.

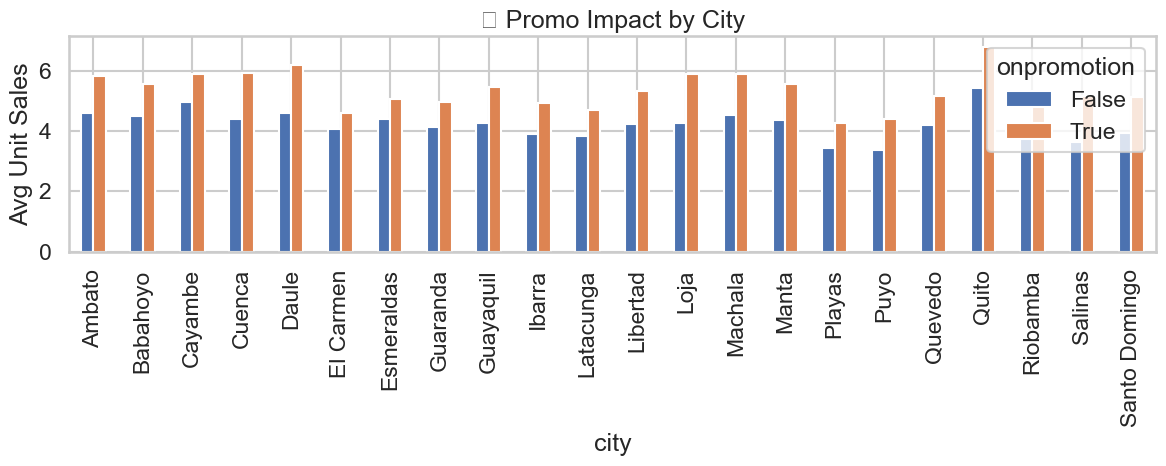
***Spatial Analysis***

The chart clearly shows that the average unit sales are significantly higher when the product is on promotion compared to when it is not. The promotion leads to an increase of roughly 1.2 units in average sales.

A graph of a number of blue and white bars

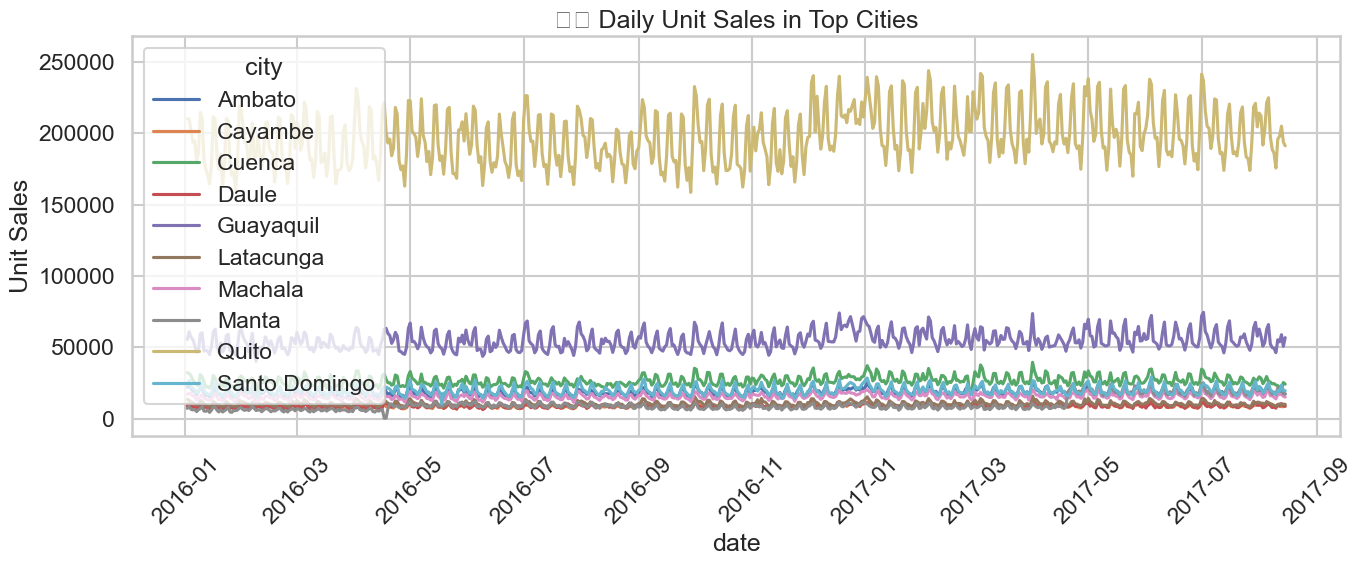
AI-generated content may be incorrect.

The chart shows that Quito has the highest average unit sales, followed by Cayambe and Daule. The average unit sales generally decrease as you move from left to right across the cities, with Puyo having the lowest average unit sales among the cities listed. There is a noticeable variation in average unit sales across the different cities.

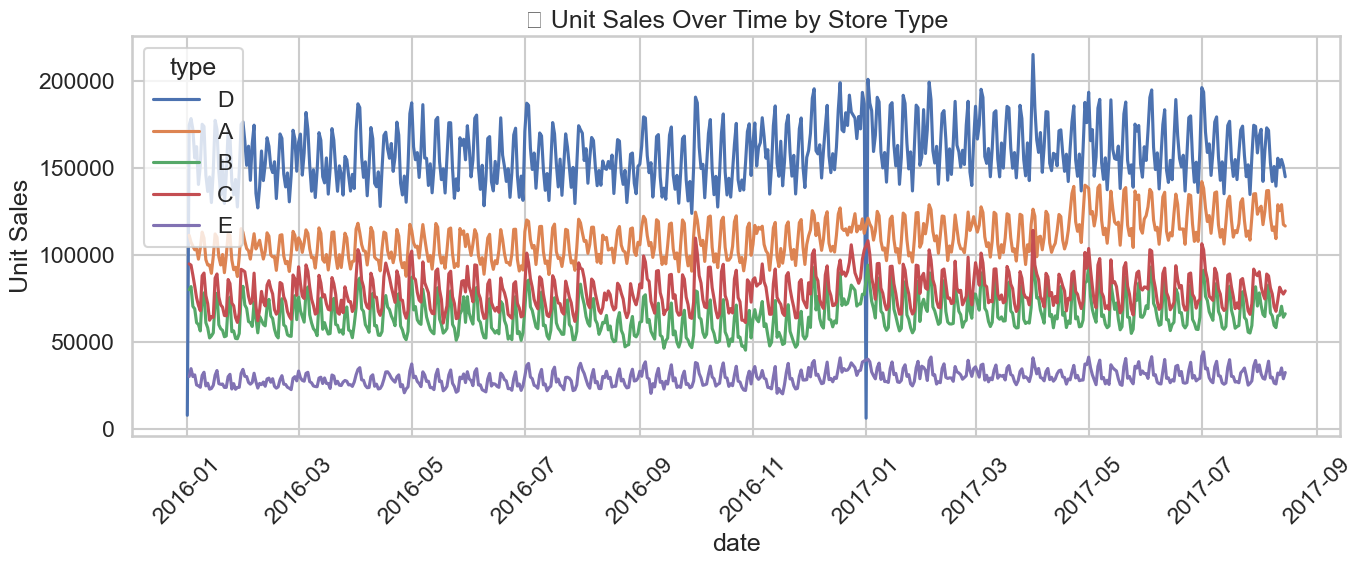


Promotions generally increase sales across cities, with the orange bar (on promotion) typically higher than the blue bar (not on promotion). The impact varies by city, with some like Babahoyo and Daule showing a larger effect, while others like Ambato and Guayaquil show smaller differences. Positive promotion impact remains consistent.

***Spatial-Temporal Analysis***



The line plot reveals that Quito consistently leads in daily unit sales, with significantly higher sales compared to other cities. Guayaquil follows with moderate sales, while the other cities show much lower, more variable sales. Seasonality is visible in Quito and Guayaquil, with recurring peaks and troughs. Notably, Quito experiences a sales dip and recovery in early 2017. The overall pattern highlights Quito and Guayaquil as clear leaders in sales performance.



The line plot depicts daily unit sales across different store types from early 2016 to mid-2017. Store type D (blue line) consistently leads in sales, significantly outperforming the other types. Store type A (orange line) follows with the second-highest sales, while store types B (green line) and C (red line) show considerably lower sales, often overlapping. Store type E (purple line) consistently has the lowest sales. All store types display a similar weekly seasonal pattern, with regular fluctuations throughout the year. A noticeable dip and recovery in early 2017 affect sales across all store types, suggesting a common event.

# **Feature Engineering**

Based on the insights obtained during exploratory data analysis (EDA), several features were engineered to capture temporal patterns, promotional effects, and regional variations in sales behavior. These features aim to enhance model performance by introducing structure and context into the dataset.

**1. Temporal Flags**

* **is\_month\_start, is\_month\_end**: Binary indicators to flag the beginning and end of each month. These were added to capture potential behavioral patterns observed during these periods, such as increased purchasing near paydays or inventory clearance at month-end.

**2. Lag Features**

* **sales\_lag\_7**: Introduced a 7-day lag in unit sales per city and product family, based on clear weekly seasonality observed in the time series plots. This helps the model learn autocorrelated patterns in sales.
* **sales\_lag\_30** *(optional)*: A 30-day lag feature to capture potential monthly cyclic trends, especially useful in cases where broader seasonal patterns exist.

**3. Seasonality Encodings**

* **day\_sin, day\_cos**: Sine and cosine transformations of the day of the week to encode weekly cyclicality without introducing artificial jumps between Sunday (6) and Monday (0).
* **month\_sin, month\_cos**: Optional sine/cosine transformations of the month variable to represent long-term seasonality throughout the year.

**4. Log Transformation**

* **log\_unit\_sales**: Applied a logarithmic transformation (log1p) to the target variable to stabilize variance and reduce skewness in the distribution. This is especially important due to the presence of outliers and high variance across categories and cities.

**5. Promotion-Based Features**

* **promo\_uplift**: Calculated as the average sales with promotion divided by average sales without promotion for each product family. This quantifies how responsive each family is to promotional campaigns, as suggested by EDA findings showing varying uplift levels across families.
* **promo\_x\_perishable**: Captures the interaction between promotional status and perishability. This was based on the hypothesis that promotions may have a stronger impact on perishable items due to urgency of consumption.
* **promo\_strength**: A city-level promotion effectiveness metric, computed as the ratio of average promotional to non-promotional sales in each city. This was inspired by observed differences in promotional impact across cities (e.g., stronger in Babahoyo and Daule, weaker in Guayaquil).

**6. City-Level Normalization**

* **city\_avg\_sales**: The average unit sales per city, used to establish a baseline for each region.
* **sales\_ratio\_vs\_city**: Represents the ratio of each row’s sales to the city’s average, allowing the model to understand relative performance across cities of varying sizes and baseline demand.

Model Building and Forecasting