# **Data Exploration Report — Store Sales (2013–2017)**

## **Overview**

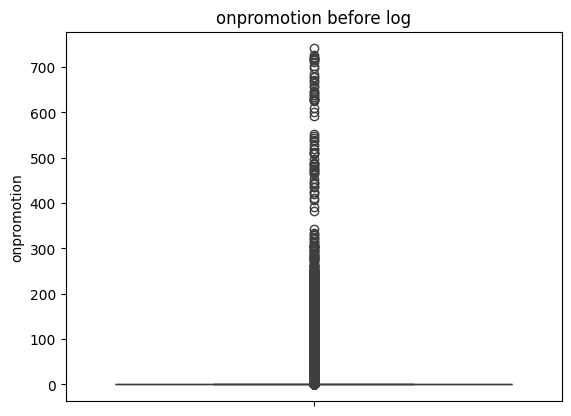
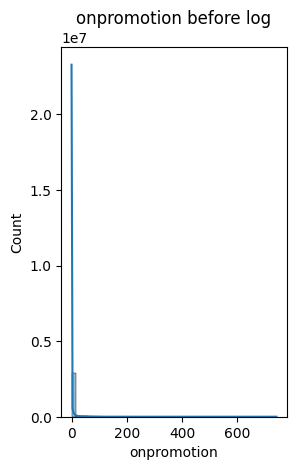
This exploration analyzes daily store-family sales from 1 January 2013 through 15 August 2017, joined with store descriptors (city, state, store\_type, cluster), holiday/event information as a categorical day\_type, and WTI oil prices (dcoilwtico). After merging and column clarification, the working training table contains 3,024,054 rows and 13 columns: date, store\_nbr, family, sales, onpromotion, city, state, store\_type, cluster, dcoilwtico, and day\_type, covering 54 stores across 22 cities and 16 states, organized into 17 clusters and 33 product families. Uniqueness checks highlighted 1,684 distinct dates; 33 families; 362 unique onpromotion values; 22 cities; 16 states; 5 store types; 17 clusters; 994 oil price points; and 6 day types. By using the figures attached to our notebook, we were able to highlight the following key points: the day-type charts (“Average Sales per Day Type,” “Max Sales per Day Type”) show Additional days lead on average while Holidays create the biggest spikes, which is helpful for planning promotions; the store-type view (“Insights from Sales by Store Type”) shows the clear ranking A > D > C > B > E; the multi-year line (“Store Sales Analysis (2013–2017)”) shows steady growth to 2016 and a dip in 2017; the monthly panel (“Monthly Sales Analysis (2013–2017)”) makes December highs, February lows, and the 2017 break (record May, sharp August drop) easy to see; the oil chart (“dcoilwtico Timeline by Year”) gives outside market context; and the transform checks (“sales before/after log,” “onpromotion before/after log”) show the log step makes the numbers more even and reduces extremes.

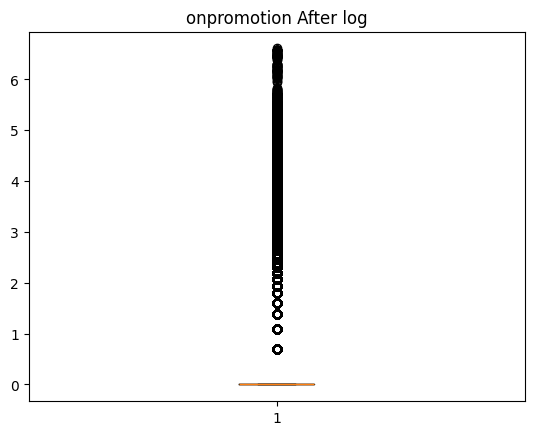
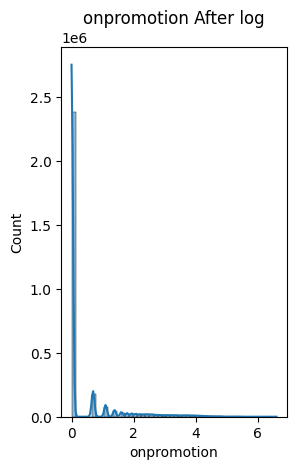
## **Data Quality and Preparation**

Before cleaning, two fields had a lot of missing data: dcoilwtico had 858,924 empty values (~28.6%), and day\_type had 2,567,862 empty values (~85.6%). We filled oil price gaps by carrying the last known value forward (and backfilled at the very start if needed). The missing day\_type was set to “Work Day.” We also cleaned text labels (trimmed spaces, used Title Case) to avoid duplicate categories. Duplicates were removed in two steps: first 30,294 rows were dropped, then a second pass removed 0 rows, producing the final table of 3,024,054 × 13. Data types are set as expected: dates for date; categories for family, city, state, store\_type, day\_type; and numbers for store\_nbr, sales, onpromotion, cluster, dcoilwtico. Sales data is right skewed: min 0.0, Q1 0.0, median 11.0, Q3 ~195.85, mean ~357.78, max 124,717.0. Oil prices range widely, between $26.19 and $110.62 (median $53.41; mean ~$67.92; SD ~$25.67). Using the IQR rule, we identified 447,105 outliers in sales and 611,329 in onpromotion (none in store\_nbr, cluster, or dcoilwtico). To make the numbers easier to model, we applied a log transform to sales and onpromotion and confirmed through the before/after plots that this reduced skew and outliers.

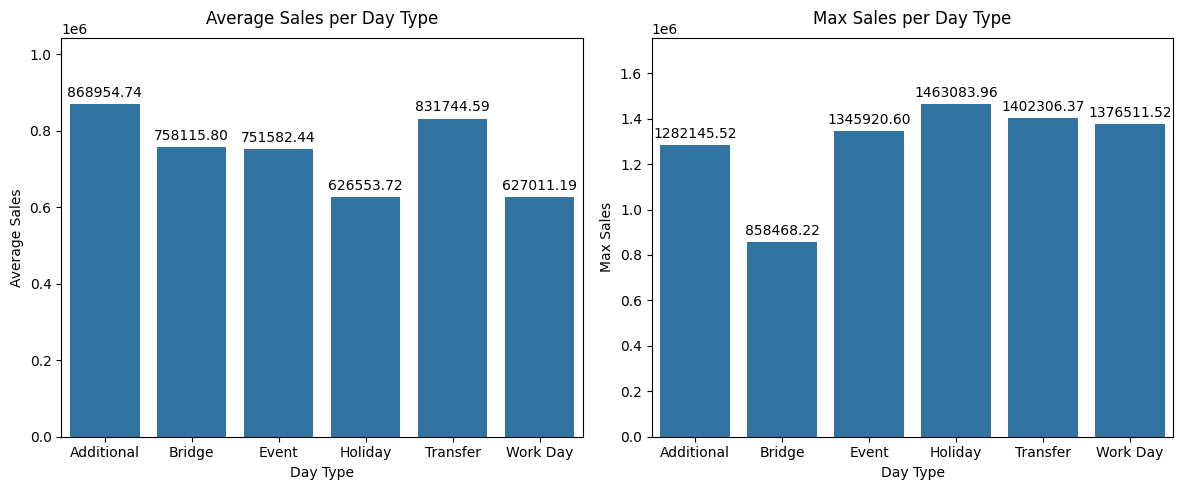
## **Engineered Features and Transformations**

We added calendar fields (Year, Month, Day, Week\_day) and a Work\_day flag from day\_type to support time-based summaries. We applied a log transform to onpromotion to reduce skew and shrink extremes; the change is visible in the existing diagnostics ( “onpromotion before/after log”). We also demonstrated one-hot encoding for categoricals and scaling for numeric fields for modeling readiness; in the encoding demo, “Work Day” was dropped as the baseline. No other features were created (no lags, rolling stats, interactions, or extra holiday flags). The log versions of onpromotion are used only for diagnostics/modeling.



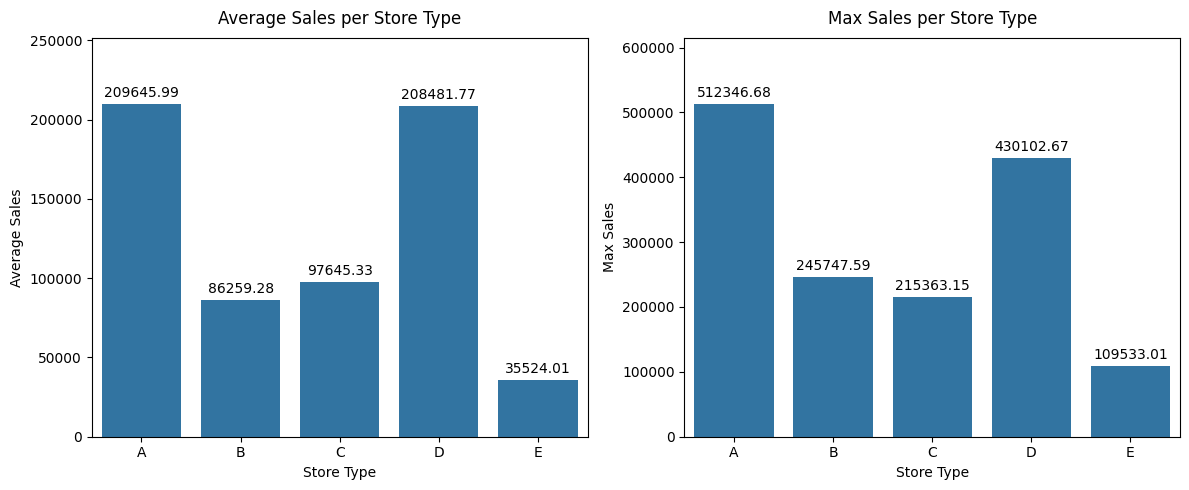


## **Trends by Day Type**

` Day type clearly affects sales. In “Average Sales per Day Type,” Additional days have the highest average (~868,955), then Transfer (~831,745), Bridge (~758,116), Event (~751,582), Work Day (~627,011), and Holiday (~626,554). In “Max Sales per Day Type,” Holiday days hit the biggest single-day highs (~1,463,084), with Transfer (~1,402,306) and Work Day (~1,376,512) close behind; then Event (~1,345,921), Additional (~1,282,146), and Bridge (~858,468). In short: Additional leads on averages, Holidays create the biggest spikes, Work Days are steady with upside, and Bridge is the weakest.

**Store-Type Performance**

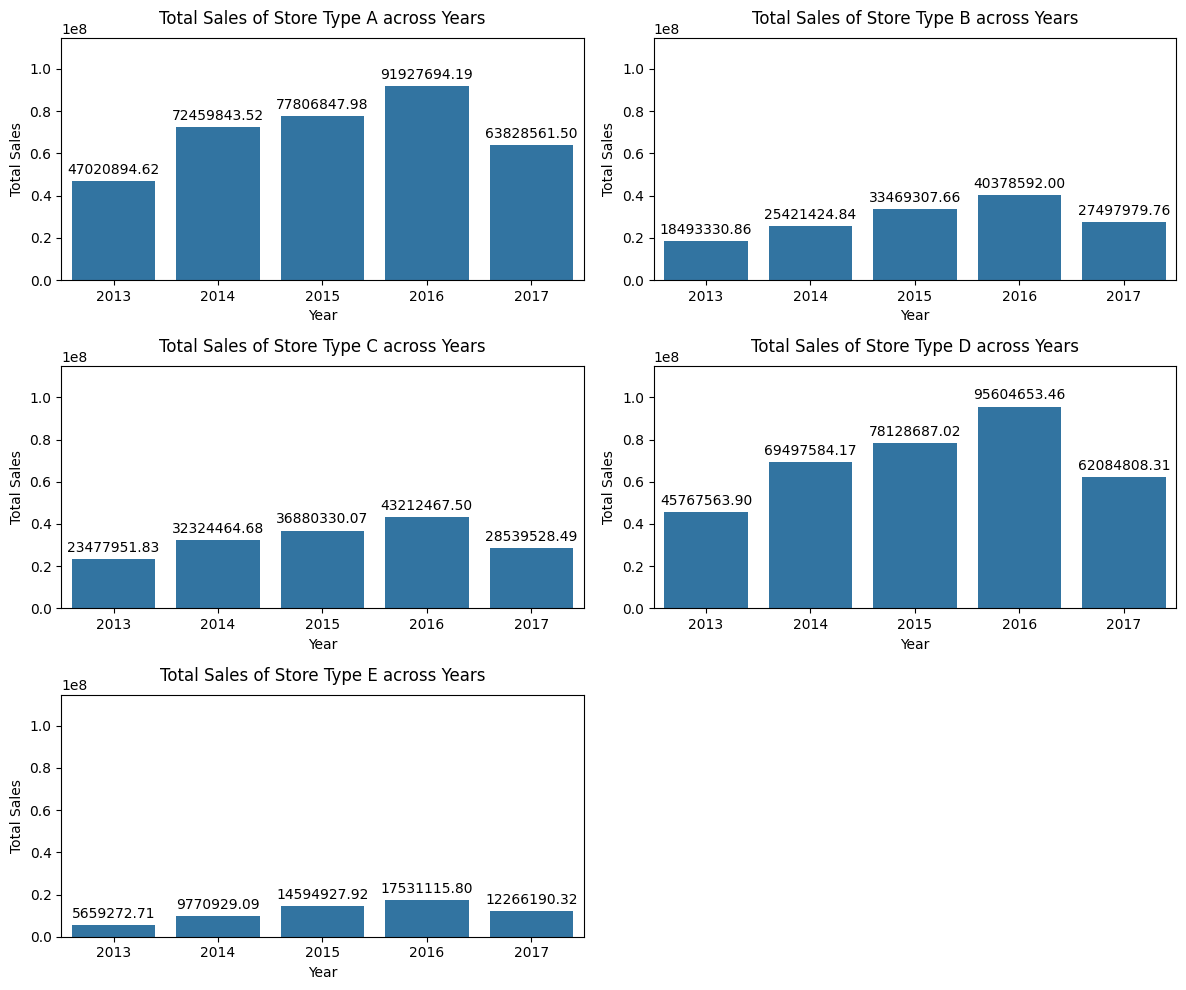
Store types show a clear, repeatable gap in performance. In “Insights from Sales by Store Type,” Type A is the top performer on both metrics with about 209,646 average sales and 512,347 max, followed closely by Type D (~208,482; ~430,103). Type C sits in the middle (~97,645; ~215,363), while Type B is lower (~86,259; ~245,748), and Type E is the weakest (~35,524; ~109,533). The ranking is identical for averages and peaks (A > D > C > B > E), so this isn’t a one-off spike; it’s a stable pattern in the data. Practically, A and D are the main volume drivers, C provides a moderate base, while B and especially E need review to understand what’s holding results back before expecting big gains. This hierarchy is useful for setting inventory levels, staffing, and promotion intensity by store type.

**

## 

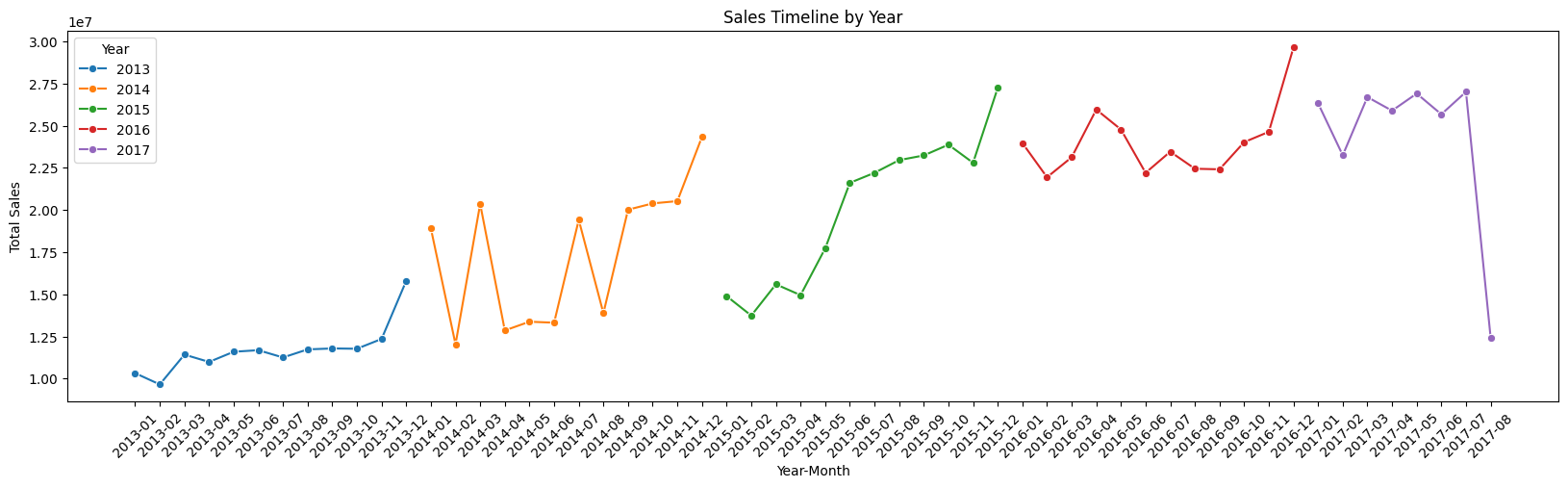
## **Store Sales Analysis (2013–2017)**

The year-by-store-type table shows the same pattern across all types: steady growth from 2013 to 2016, then a drop in 2017. For example, Type D totals are 45,767,564 (2013), 95,604,653 (2016), and 62,084,808 (2017). Type A is 47,028,945; 91,927,694; and 63,828,562. Type C is 23,477,952; 43,212,468; and 28,539,528. Type B is 18,493,331; 40,378,592; and 27,499,780. Type E is 5,659,273; 17,531,116; and 12,266,190. A summary view in the notebook also marks 2016 as the peak year for every type (e.g., D ≈ 95.6M, A ≈ 91.9M) and 2013 as the lowest (e.g., D ≈ 45.8M, A ≈ 47.0M). The “Store Sales Analysis (2013–2017)” chart makes this rise through 2016 and the broad pullback in 2017 clear.



## **Seasonality**

Monthly results show a clear, steady pattern. From 2013 to 2016, total sales grow from about 10–12M per month to ~25–30M. December is the top month each year (≈15.9M in 2013; 24.4M in 2014; 27.3M in 2015; 29.8M in 2016), and February is the lowest (≈9.8M in 2013; 12.0M in 2014; 13.8M in 2015; 22.0M in 2016). This makes the seasonal shape predictable: a strong year-end lift and a softer early-year dip. In 2017, that shape changes: May hits a new high (~27.0M), followed by a sharp August drop (~12.5M). This differs from earlier years where December led. The notebook also notes that late-2017 months may be missing, so results should be checked against the usual September–December bump once data is confirmed.



## **Conclusion and Key Findings**

From 2013–2016, sales exhibit stable seasonality—consistent December peaks and February troughs—followed by a clear regime change in 2017, when May sets a new high and a sharp summer pullback appears. By day type, Additional days lead on average, Holidays generate the largest single-day spikes, Work Days are steady with room to grow, and Bridge days underperform. By store type, the hierarchy is stable at A > D > C > B > E, identifying A and D as the principal volume drivers and E as the weakest. Data quality issues were concentrated in auxiliary fields; we forward-filled (and early backfilled) missing oil prices, defaulted missing day\_type to “Work Day,” removed duplicates, and standardized labels. Because sales and onpromotion are heavily right-skewed, a simple log transform improves distributional shape for modeling. Overall, the EDA provides a clean, model-ready base and highlights two priorities for forecasting: (1) accommodate the 2017 break in seasonality, and (2) focus on weak segments, Bridge days and store types B and especially E, during feature design and model evaluation.